



ISSN 2282-6483

Alma Mater Studiorum - Università di Bologna
DEPARTMENT OF ECONOMICS

**Building a firm level dataset for the
analysis of industrial dynamics and
demography**

Marco Grazzi
Chiara Piccardo
Cecilia Vergari

Quaderni - Working Paper DSE N°2003



BUILDING A FIRM LEVEL DATASET FOR THE ANALYSIS OF INDUSTRIAL DYNAMICS AND DEMOGRAPHY*

Marco Grazzi¹, Chiara Piccardo¹, and Cecilia Vergari¹

¹Department of Economics, University of Bologna

May 16, 2017

Abstract

This paper illustrates the building procedure of a firm-level panel dataset that merges several sources of information concerning the various activities of business firms. The aim of this work is to achieve a detailed dataset able to shed light on firm demographics, in terms of survival, entry and exit processes, distinguishing between “voluntary” and “involuntary” exits. Moreover, the derived dataset allows to monitor the innovation activities of the firms and also to capture complementarities between two instruments of intellectual property rights (IPRs), namely granted patents and registered trademarks. We assess the validity of the proposed procedures resorting to the virtual universe of Italian limited liability companies as provided by Bureau van Dijk (BvD). The dataset covers more than 1 million companies operating in both manufacturing and service sectors and contain financial and economic information, as well as, among the others, the ownership structure and administrative procedures undergone by the firms, which may lead to firm exit. The main purpose of the paper is to provide a unified set of procedures to help the researcher dealing with the vast amount of information available on corporate firms and of ever increasing size. This will also facilitate the replication of empirical analyses, across researchers working on dataset with similar characteristics, although from different countries or data providers.

Keywords: Firm-level data, firm demography, integrated database, Structural Business Statistics, patents, trademarks

JEL classification: C81, L60, L80, O14, O34

*We thank the participants to the “Organizzazione e strategie innovative d’impresa in economia aperta” workshop in Forlì, June 2016. We also acknowledge useful remarks by Roberta Rolando on the definition of firms exit. Moreover, we are grateful to Roberto Susanna (Infocamere press office) and to Bureau van Dijk for their technical assistance at various stage. We are also indebted to Daniele Moschella for insightful comments on IPR data management. All remaining errors are our own. This work was completed within the research grant “Organizzazione d’impresa, innovazione e commercio internazionale” founded by the Fondazione Cassa di Risparmio of Forlì. Corresponding author: Chiara Piccardo, chiara.piccardo@unibo.it.

1 Introduction

The last two decades witnessed the increasing availability of a vast amount of data that became available to researchers in almost all fields in the social sciences. Economics was no exception as it could take advantage of many disaggregated, individual level dataset, where the unit of observation can be, among the others, the consumer, the household or the firm. It was certainly the firm-level dataset that registered one of the most significant surge, probably because the issues related to data sensitivity and anonimization were easier to be taken care of.

Since the seminal work of Baily et al. (1992), one of the first to employ firm level data, many others took advantage of this disaggregated source of data to investigate, for instance, employment, productivity, demography, as well as, the innovation and export activities of firms (see, among the many others, Baldwin and Rafiquzzaman, 1995; Davis et al., 1996; Bartelsman and Doms, 2000).

The rising availability of firm-level data allowed the attainment of numerous developments in the discipline. One of the most prominent was certainly highlighting the wide and persistent heterogeneity existing across firms operating within the same sector of economic activity. This is well documented by a large body of research from different industries and countries (cf. Baily et al., 1992; Bartelsman and Doms, 2000; Disney et al., 2003; Syverson, 2011; Dosi et al., 2016, among many others) which point to the emergence of a few “stylized facts”: wide asymmetries in productivity across firms; significant heterogeneity in relative input intensities; high intertemporal persistence in the above properties and, finally, the fact that such heterogeneity is maintained also when increasing the level of disaggregation.

This latter property, to which the availability of disaggregated, firm-level data greatly contributed, was sharply put forth by Griliches and Mairesse (1999): “*We [...] thought that one could reduce heterogeneity by going down from general mixtures as “total manufacturing” to something more coherent, such as “petroleum refining” or “the manufacture of cement.” But something like Mandelbrot’s fractal phenomenon seems to be at work here also: the observed variability-heterogeneity does not really decline as we cut our data finer and finer. There is a sense in which different bakeries are just as much different from each others as the steel industry is from the machinery industry.*”

Notwithstanding, as recalled above, all the improvements brought to the discipline by the use of firm-level data there is still a relative under-exploitation of disaggregated source of data to uncover basic relations at the micro level, among the variables of interest. Or to put it in other terms, too much is still assumed from theory and left untested. Part of the explanation, we claim, is due to the hurdles and complications that often dissuade the researcher from engaging in empirical analysis with firm-level data. The data collected by National Statistical Offices or other institutions are indeed far from being ready-to-use and require a considerable investment of time and a wide range of competences including - but not limited to - the thorough understanding of industrial and product classification, a basic understanding of firm and employment regulation and a good command of data management and programming skills. In this work, we describe a series of procedures that enable, starting from “raw” firm level data, to attain a dataset that can be employed for empirical analysis (for previous works describing the development of firm-level datasets see, among others, Grazzi et al., 2013; Kalemli-Ozcan et al., 2015). In this respect, we show how to overcome a series of problems that often arise, such as, merging together dataset with different levels of observation (i.e. firms *versus* patents), determining the

“proper” entry or exit of a firm from the dataset, and others.

Another goal that we aim at with this work is contributing to replicability of empirical analyses in the social science. It is well known, indeed, that the replication of empirical work is highly costly in terms of time or compromised by different “cleaning” procedures applied on the same set of data by different researchers. In this respect, contributing to establishing a common set of rules, will make it easier to replicate results.

In what follows, we will apply this set of procedures to a firm-level dataset of Italian companies, AIDA, provided by Bureau van Dijk, BvD henceforth. However most of the procedures described can be applied to other firm level dataset from other countries or provided by other companies or institutions. In particular, we first describe the dataset AIDA and the extracting procedure we followed (Section 2). We next focus on the steps needed to bring the dataset to the standard panel data format for accurate analysis of firms’ demography (Section 3). In Section 4, we describe the coverage of AIDA as compared with official data from Eurostat Structural Business Statistics and INFOCAMERE. Section 5 reports the procedure to distinguish between “voluntary” and “involuntary” exit. Finally, Section 6 illustrates the merging procedure of AIDA with information on firms’ intellectual property rights. We conclude in Section 7.

2 Accessing and extracting AIDA data

The AIDA dataset contains detailed information on Italian limited liability firms as they are required to deposit the balance sheet to the local Chamber of Commerce. BvD then assemble the data. The user can access AIDA data by two alternative ways: BvD proprietary browser online (aida.bvdinfo.com) and BvD physical media (CD-ROM/DVD, Blu-Ray). Each of these two methods has advantages and disadvantages for the researcher; however, to the best of our knowledge, the latter is the best choice as it is less time consuming in case of academic research requiring large volumes of data.¹ For this work we used the AIDA DVD (December 2015) that covers the period between 2005 and 2014.

AIDA reports financial-economic information on the virtual universe of limited liability companies operating in Italy.² In order to facilitate the research of relevant information, the data are organized in ten sections: identification number, contact details, legal and account information, account header, size and group information, industry overview, financial and ratios, stock data, directors/managers/contacts and auditors, ownership data. In order to quickly and easily analyze firms data, the researcher can identify and save a list of variables of interest. This chance turns out to be particularly useful in case of research developed at different times; indeed, once the list of variables has been saved, the user can import the selected variables avoiding wastes of time and omissions due to researcher’s forgetfulness.

The list of variables we focused upon is reported in Tables 1 and 2. In particular, Table 1 includes the list of “static variables”, i.e. variables that are available only in the last year of the dataset (2014). The majority of these variables, with the exception for *legal*

¹Accessing the data from aida.bvdinfo.com imposes stringent limits on the number of firms and variables user can extract in one run and, generally, the extracting process is slow. On the contrary, these limits do not apply for retrieving the data from AIDA-DVDs.

²All limited liability companies have to deposit their balance sheet, however as in all firm-level dataset there are missing and there is some attrition. Hence the expression “virtual universe”.

Table 1: List of static variables

AIDA section	Variables of interest
Identifiation numbers	VAT number
	BvD ID number
Contact details	Municipal ISTAT code
	Province ISTAT code
	Region ISTAT code
Legal and account information	Previuos CCIAA
	CCIAA change date
	Legal status
	Legal form
	Date of incorporation
	Last accounting closing year
	Pending administrative procedures
	Beginning of administrative procedure
	End of administrative procedure
Accounts header	Consolidation code
Size and group information	BvD independent indicator
	Number of companies in corporate group
	Number of recorded shareholders
	Number of recorded subsidiaries
Industry and ownership	ATECO 2007 code
	NACE Rev. 2 code
Directors and managers contacts	Number of current directors, managers contacts
Ownership data (immediate parent information)	BvD ID number
	Country ISO code
	NACE Rev. 2, Core code

Note. All variables are reported as “static”, meaning that their values are referred to the last available year (2014) in the AIDA-DVD (2015).

*status, pending administrative procedures (procedure/cessazione), beginning of administrative procedure (date of opening of the procedure) and end of administrative procedure (date of closure procedure/cessazione),*³ is not expected to vary over the lifetime of the firm; therefore, we have considered them as constant over the period under observation. These information include, among others, firms’ identification number, registered office address, legal form, year of incorporation, corporate group items, industry and ownership structure.

Table 2 shows the list of financial variables. The AIDA-DVD (2015) provides the last ten years of balance sheet data for the same companies, thus these variables are referred to the period 2005-2014.⁴

Before extracting data, the user has to choose the time span for the financial and economic information. The researcher has to choose between two alternative options: “absolute” and “relative” years, respectively. If choosing “absolute” years, the user has to specify the calendar years (e.g. 2005, 2006 and so on). When choosing “relative” years, the researcher has to select the latest available years for financial and economic data (e.g. *last avail. year* identifies the most recent available non-missing data, while

³We postpone discussion about these variables to Section 5 where we define firms’ entry and exit from the market.

⁴Note, however, that, preliminary exploratory work revealed that in most cases there is a reporting lag of about two years; hence, it is safe to assume that in the AIDA-DVD (2015) the last reliable year for our purpose is 2013.

Table 2: List of financial variables

AIDA section	Variables of interest
Financial and ratios	Total shareholder's funds
	Number of employees
	Total fixed assets
	Total intangible fixed assets
	R&D expenditure
	Industrial patents and intellectual property rights
	Concessions, licenses, trademarks and similar rights
	Total tangible assets
	Total financial fixed assets
	Total current assets
	Total assets
	Total payables
	Due to bank
	Due to bank - beyond 12 months
	Due to other lenders
	Due to other lenders - beyond 12 months
	Due to suppliers
	Due to suppliers - beyond 12 months
	Total value of production
	Revenues from sales and services
	Raw, consumption materials and goods for resale
	Services
	Total personnel costs
	Wages and salaries
	Total depreciation, amortization and writedowns
	Operating margin
	Added value
	Profit and loss before taxation
	Liquidity ratio
	Current liabilities/total assets
	Long and medium term liabilities/ total assets
	Leverage
	Cost of debit
	Solvency ratio
	EBITDA
	Return on sales - ROS
	Gross profit
	Cash flow

Note. All variables are referred to the period 2005-2014.

last avail. year-1, last avail. year-2, and so on refer to the earlier available non-missing information). Even if for companies with gaps in their data the option “relative” years might cover a more extended time span, we chose to extract financial and economic data selecting the “absolute” years for the period 2005-2014. Other choices that the user has to make before extracting the data regard units and currency. In our case, we selected, as default option for AIDA, thousands as unit of measure and Euro as currency.⁵

The group of companies under investigation can be identified in AIDA by employing several criteria (e.g. location, industry, legal form and number of employees, among others) and the user can combine them by using full Boolean logic (and, or, and not, from). In order to obtain the most comprehensive group of companies, we opted for considering the location as selection criteria; thus, we included in the dataset only firms for which the Italian region (corresponding to company’s address) was not missing. We obtained a total of 1,298,919 firms, which roughly corresponds to the universe of limited liability companies in Italy. Note that the large size of the dataset and the almost complete representativeness of the population is crucial for the analysis of firms demographics. Generally, new firms are characterized by small size and do not have any employees, thus, the dataset virtually identifies real firms entry. By contrast, using a dataset with a threshold on employment might not capture the effective entries.

Once the set of variables of interest has been identified and the firms have been selected, the user can visualize and/or export the resulting list of companies. In particular, AIDA offers a variety of export formats for data: excel, text and XLM. We exported data to a tab delimited text file, which corresponds to the most appropriate export format for a large amount of data; the drawback of this format is that data need to be further processed in order to be usable.⁶

3 Data from “wide” to “long”

We extracted data from AIDA, selecting the period between 2005 and 2014 for the financial variables, in tab separated text format.⁷ We imported the file in STATA specifying the UTF-16 encoding option.⁸ The data came in STATA “wide” format, meaning one row for each company and several columns for each variable-year combination. As a matter of illustration a row incorporated BvD ID (identification) number, “static” variables (time invariant data) and financial variables (time varying information), identified by the variable’s name followed by the year of reference (e.g., ValueAdded2014, ValueAdded2013, ..., ValueAdded2005). To bring the data in the more standard “long” format, where for each firm each year of data is in a separate row, it is necessary to “reshape” the dataset. In the following, we describe the procedures needed before running the actual “reshape” command.

⁵The AIDA dataset reports either firms’ consolidated or unconsolidated balance sheet data. However, for some companies AIDA provides both types of data. For these companies we only extracted data from the unconsolidated balance sheet.

⁶Tab delimited text files are encoded in UTF-16, which means that characters are represented as binary sequences. Each sequence is made up by either one or two 16-bit integers.

⁷It is possible to speed up the exporting process by downloading separated files, each including firms located in a subset of Italian regions.

⁸We employed Stata (version 14) as it is one of the most widely diffused software in economics and social sciences. The same procedures can be replicated also with other non-proprietary software such as R.

In AIDA, companies are uniquely identified by the BvD ID number;⁹ in principle, the identification number does not change over time.¹⁰ Some “static” variables, such as *previous CCIAA*, *CCIAA change date*, *pending administrative procedures*, *beginning of administrative procedure* and *end of administrative procedure*, take more than one value for some companies, thus for these companies we obtained as many rows as the number of values taken by the variable of interest and we needed to replicate the BvD ID in each of these rows. Moreover, we renamed variables and harmonized their format to avoid technical problems related to conflicts in variables storage types, and we removed the BvD ID which had been double counted in the files.

The set of procedures described here aims at building a dataset that allows the accurate analysis of firms demography.¹¹ To preserve relevant information on variables regarding the administrative procedures that are associated to firm exit, it was necessary to proceed as follows. To start, we converted the string variable *pending administrative procedures* to numeric format¹² and, then, focused on administrative procedures undergone by firms between 2004 and 2014.¹³

Information on administrative procedures undergone by firms (*pending administrative procedures*) and the related *beginning of administrative procedure* were repeated in different rows for each firm, thus they were in “long” format. We made these variables uniform to the rest of the dataset which was in “wide” format. In our sample, the maximum number of administrative procedures undergone by a firm in any given year was five.¹⁴ Considering the 11 years making up the dataset (2004-2014), we generated 55 variables reporting the administrative procedures undergone by a firm in chronological order (e.g. PROC1_2004 reported the first administrative procedure undergone by a firm in the year 2004, while PROC5_2004 provided the last one undertaken by the firm in the same year). Similarly, we generated 55 variables related to the *beginning of administrative procedure*; where, for instance, DATE1_2004 reported the date for the first administrative procedure undergone by a firm in the year 2004, while DATE5_2004 provided the date for the last one undertaken by the firm in the same year.

After the creation of these variables, the dataset turned out to be in a uniform wide format so that it was possible to convert it to “long” without loss of information.¹⁵

So far, we have explained how we accessed, extracted and organized the data in the “long” format by resorting to the data stored on a single physical DVD. In the reminder

⁹The BvD ID number allows the user to obtain data from other BvD products about the relevant group of firms, provided that these products cover the companies in question.

¹⁰In a very limited number of cases these identifiers could vary through time and a new BvD ID number can be assigned to a company. These changes occur as consequence of firms changes of address, legal form, or merger and acquisition activity. Nevertheless, the number of companies experiencing a BvD ID change is negligible when compared to the bulk of companies included in the dataset.

¹¹We postpone to Section 5 for a more detailed explanation on the construction of the firms’ exit indicator.

¹²It has been necessary in order to easily handle the variable. We did not use the “destring” command available in STATA 14, but we associated to each administrative procedure an integer number. The generated variable assumes values from 1 to 66.

¹³We considered administrative procedures with the *beginning of administrative procedure* between the 1st January 2004 and the 31st December 2014. Financial information are available for the period 2005-2014.

¹⁴We do acknowledge that we were ourselves rather surprised to find firms reporting more than one extra-ordinary event affecting their “administrative” life in a given year. This contributes to lend support to the perception of Italy as a country with a relatively high administrative burden for firms.

¹⁵In order to speed up the reshape process, it is possible to split the dataset in more than one file and drop some string variables which made the data manipulation more demanding.

of this section we provide some general guidelines combining data from different disks. Each AIDA DVD provides ten years of balance sheet data; thus, the researcher might obtain a longer time span by extracting data from more than one disk.

For the sake of consistency, the user should select firms from different disks by using the same selection criteria and identifying the same set of variables of interest. In our case, we simply repeated the procedures that we described in Section 2.

Once extracted all the relevant data from different disks, before merging datasets, one shall convert each of them from “wide” to “long”. In our case, in order to get uniform datasets we should follow all the procedures illustrated in Section 3 for data from each disk.

Finally, the user should merge all the available datasets by using the unique firm ID.

4 Coverage of the AIDA dataset

Once the dataset is in the more standar panel data format, it is possible to investigate its coverage. In particular, in this Section we propose a comparison between the AIDA dataset and official data from Eurostat Structural Business Statistics (EUROSTAT SBS) and INFOCAMERE (Movimprese).¹⁶

We start by focusing on the comparison between AIDA and EUROSTAT SBS data in terms of the coverage of the number of firms in each year (2005-2014), as showed in Table 3. EUROSTAT SBS data are available for the period 2005-2014 for economic sectors from Sections B to N and Division S95 of NACE Rev.2¹⁷ and, for each year, they include the population of active enterprises irrespectively of their legal form. Thus, for the sake of consistency, we only considered firms included in the AIDA dataset operating in economic sectors covered by the EUROSTAT SBS data.¹⁸ In the period under observation the coverage of the AIDA dataset ranges between 13.12% and 20.13%. Note that this apparent under-representation is almost completely due to the fact that, while EUROSTAT SBS includes limited and unlimited liability companies as well as personally owned firms, AIDA only contains limited liability companies. In terms of representativeness of economic activities both datasets assign a similar share of firms to manufacturing and not-manufacturing sectors (less than 20% of firms operates in manufacturing sectors and more than 80% in service sectors in both data, respectively).

Table 4 shows the size distribution of firms in AIDA compared with EUROSTAT SBS data. We considered three firms’ size classes: “small” includes firms with the number of employees ranging between 0 and 19; “medium” comprises firms with number of employees ranging between 20 and 249 and “large” includes firms with more than 250 employees.¹⁹ Both sources of data highlight a very asymmetric size distribution, with a higher fraction of Italian firms classified as small, while just a lower fraction defined

¹⁶INFOCAMERE is the company of the Italian Chambers of Commerce that takes care of processing the data coming from the balance sheet of limited liability firms. Movimprese is a report on firms’ death and birth provided by INFOCAMERE every quarter.

¹⁷EUROSTAT SBS partially covers Section K of NACE Rev.2 on insurance services, credit institutions and pension funds. These data are not available for Italy.

¹⁸In order to perform the comparison between AIDA and EUROSTAT SBS, we only considered the sectors with available information in EUROSTAT SBS dataset. Thus, for the period 2008-2014 we considered the Nace Rev. 2 Sections form B to N (excluding K) and Subsection S95, while, for period 2005-2007, we did not include even the Nace Rev. 2 Section E and M.

¹⁹Number of employees is defined as those persons who work for a firms and who have a contract of employment and receive compensation in the form of wages, salaries, fees, gratuities, piecework pay.

Table 3: Coverage of the AIDA dataset relative to EUROSTAT SBS and sector distribution

year	All firms		Manufacturing		Non-Manufacturing	
	(EU SBS)	(% AIDA)	(% EU SBS)	(% AIDA)	(% EU SBS)	(% AIDA)
2005	3,227,588	13.12	14.93	21.89	85.07	78.11
2006	3,235,790	14.09	14.73	21.22	85.27	78.78
2007	3,257,467	15.02	14.51	20.61	85.49	79.39
2008	3,948,726	14.41	11.64	18.38	88.36	81.62
2009	3,889,543	15.49	11.29	17.92	88.71	82.08
2010	3,867,813	16.16	11.03	17.53	88.97	82.47
2011	3,843,454	16.81	11.07	17.20	88.93	82.80
2012	3,825,458	17.46	10.91	16.95	89.09	83.05
2013	3,770,844	18.67	10.80	16.69	89.20	83.31
2014	3,715,164	20.13	10.67	16.38	89.33	83.62

Notes. Column II shows the number of firms in EUROSTAT SBS; column III reports the percentage coverage in AIDA; columns IV and V exhibit the share of firms in manufacturing sector in EUROSTAT SBS and in AIDA, respectively; columns VI and VII display the share of firms in non-manufacturing sector in EUROSTAT SBS and in AIDA, respectively.

as large. Note that the smaller size bias apparent from the EUROSTAT SBS figure is mostly due to the presence of unlimited liability firms in EUROSTAT SBS data.

A more appropriate comparison to assess the representativeness of the database is possible by resorting to INFOCAMERE data which allow to select firms according to their legal form. INFOCAMERE data also provides information on all economic sectors and on the geographical distribution. We considered the period between 2005 and 2014 and, in order to have a more accurate comparison, we only included joint stock companies, limited partnerships with shares and limited liability companies.²⁰ As reported in Table 5, firms included in AIDA represent around 75% of the population according to INFOCAMERE dataset.²¹ As suggested by both INFOCAMERE press-office and BvD division, the difference on the coverage of the two datasets is mainly due to the fact that AIDA only includes firms registered to the register of companies that actually deposit their balance sheet to the Italian Chambers of Commerce. On the other hand, INFOCAMERE dataset covers all firms showing up in the register of companies irrespectively of their status and of whether or not they delivered their balance sheet.²² Moreover, as we will explain in the next section, we excluded from our dataset those companies that have not undergone any administrative procedure and have been deleted from the AIDA dataset because they did not report their balance sheet in the last 5 years. In terms of sector distribution, it was possible to distinguish between firms operating in primary, manufacturing and service sectors.²³ Table 5 suggests that the sectoral distribution of firms in the AIDA

²⁰We did not account for other legal forms included in AIDA, such as associations, consortium and cooperative companies, among others.

²¹In order to identify the coverage of the AIDA dataset with respect to INFOCAMERE data we accounted even for firms which did not provide information on their economic sector.

²²INFOCAMERE data include all firms appearing in the register of companies that filed the “certified notification of setting up of business”.

²³For the period 2005-2008, we did not consider firms in the ATECO 2007 Sections O and U; these Sections are not covered by the INFOCAMERE data for these years. For the period 2009-2014, we accounted for all ATECO 2007 Sections. The ATECO 2007 classification of economic activities is the Italian national version of the European classification (Nace Rev.2).

Table 4: EUROSTAT SBS and AIDA datasets: size distribution

year	0 to 19 employees		20 to 249 employees		250 + employees	
	(% EU SBST)	(% AIDA)	(% EU SBS)	(% AIDA)	(%EU SBS)	(%AIDA)
2005	98.41	78.57	1.55	19.96	0.04	1.47
2006	98.40	82.12	1.56	16.83	0.05	1.05
2007	98.36	88.66	1.60	10.66	0.05	0.68
2008	97.99	92.32	1.93	7.17	0.08	0.51
2009	98.07	93.18	1.85	6.35	0.08	0.48
2010	98.09	93.48	1.82	6.05	0.08	0.47
2011	98.11	90.99	1.81	8.55	0.08	0.46
2012	98.14	91.42	1.78	8.14	0.08	0.44
2013	98.20	91.92	1.73	7.66	0.08	0.42
2014	98.19	92.47	1.73	7.14	0.08	0.39

Notes. Each cell corresponds to the share of firms in the indicated size class with respect to the total number of firms from the EUROSTAT SBS and AIDA data, respectively.

dataset turns out to be quite similar to firms included in the INFOCAMERE data. In particular, in both sources, about 80% of firms operates in services sectors, about 20% in manufacturing sectors and only a lower fraction of companies operates in the primary sectors (slightly more than 1%). Thus, we can conclude that AIDA well represents the “true” distribution of Italian firms’ across sectors.

Table 6 proposes a comparison between AIDA and INFOCAMERE data in terms of firms’ size distribution. In particular, as INFOCAMERE dataset provides a more appropriate reference point for AIDA, we now employ a more fine-grained distribution of firm size classes and consider micro (0-9 employees), small (10-49 employees), medium (50-249 employees) and large firms (more than 250 employees).²⁴ While we referred to the number of employees to define size classes, in INFOCAMERE data size classes are identified by the number of workers, which includes not only employees but even self-employed workers.²⁵ Table 6 shows that the highest share of firms, in both data, are micro-firms; in particular, this category of firms is mainly made up by firms with a number of employees lower than or equal to 5 units (about 75% of firms in both AIDA and INFOCAMERE data).²⁶ The share of small firms is about 12% of firms and the fraction of medium firms is about 2%, while the lowest portion of firms is classified as large (less than 0.5% in both datasets).²⁷

Finally, Table 7 shows that most firms are located in the North-West of Italy (more than 30% on average over the entire period considered) and slightly more than 20% on average over the entire period are located in the North-East and Center of Italy, respectively. Instead, the lowest share of firms is located in the Isles (about 5%).²⁸

²⁴ISTAT (Italian National Institute for Statistics) suggests this firms’ size classification (ISTAT, 2015).

²⁵Data on number of workers are not available at the firm level for the period 2005-2007 in INFOCAMERE data, thus we limited the firms’ size distribution comparison to the period 2008-2014.

²⁶The size class 0-5 employees includes firms without information on workers (INFOCAMERE data) and employees (AIDA).

²⁷Both sources of data also include firms without information on their economic sector.

²⁸North-West includes firms located in Liguria, Lombardy, Piedmont and Valle D’Aosta; North-East comprises firms located in Emilia-Romagna, Friuli-Venezia Giulia, Trentino-Alto Adige and Veneto; Centre includes firms located in The Marches, Tuscany, Lazio and Umbria. Isles covers firms located in Sardinia and Sicily, while South covers firms located in the remaining regions. The table include firms without information on their economic sector from both INFOCAMERE and AIDA datasets.

Table 5: Coverage of the AIDA dataset relative to INFOCAMERE data and sector distribution

Year	All firms		Primary		Manufact.		Service	
	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)
2005	670953	77.16	1.42	1.36	20.33	19.10	78.25	79.55
2006	710445	78.63	1.41	1.34	19.69	18.45	78.90	80.21
2007	755187	79.27	1.40	1.33	19.02	17.86	79.58	80.80
2008	878005	72.46	1.38	1.37	18.23	17.35	80.38	81.28
2009	903666	73.97	1.33	1.40	16.81	16.86	81.87	81.74
2010	929340	72.50	1.38	1.47	16.52	16.52	82.11	82.01
2011	953949	70.50	1.42	1.51	16.16	16.30	82.41	82.18
2012	966141	69.51	1.47	1.54	15.87	16.15	82.66	82.31
2013	982943	69.58	1.48	1.54	15.60	16.02	82.92	82.44
2014	1008451	71.57	1.49	1.51	15.35	15.71	83.16	82.78

Notes. Column I shows the number of active limited liability companies in INFOCAMERE data; column II reports the percentage coverage in AIDA; columns III and IV exhibit the share of firms in primary sectors in INFOCAMERE dataset and AIDA, respectively; columns V and VI show the share of firms in manufacturing sector in INFOCAMERE data and in AIDA, respectively; columns VII and VIII display the share of firms in non-manufacturing sector in INFOCAMERE dataset and in AIDA, respectively.

Table 6: INFOCAMERE and AIDA datasets: size distribution

year	0-5		6-9		10-19		20-49		50-99		100-249		250-449		500 +	
	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)	(IX)	(X)	(XI)	(XII)	(XIII)	(XIV)	(XV)	(XVI)
2008	76.21	78.42	8.25	7.17	8.40	8.26	4.73	3.52	1.34	1.42	0.73	0.81	0.20	0.23	0.15	0.17
2009	74.82	82.23	8.96	5.87	8.96	6.36	4.86	3.07	1.34	1.31	0.71	0.77	0.20	0.22	0.15	0.16
2010	75.24	85.36	9.04	4.35	8.76	5.02	4.65	2.88	1.29	1.27	0.68	0.74	0.19	0.22	0.14	0.16
2011	74.28	73.50	9.52	9.16	9.09	9.48	4.78	5.19	1.32	1.50	0.68	0.79	0.19	0.22	0.14	0.16
2012	74.44	72.77	9.58	9.60	9.02	9.76	4.68	5.22	1.29	1.48	0.67	0.78	0.19	0.23	0.14	0.16
2013	75.41	73.55	9.37	9.48	8.64	9.42	4.42	4.96	1.21	1.43	0.64	0.77	0.18	0.22	0.13	0.16
2014	76.08	73.81	9.18	9.57	8.37	9.29	4.26	4.82	1.18	1.39	0.63	0.75	0.17	0.21	0.13	0.16

Notes. Each cell corresponds to the share of firms in the specified size class respect to the total number of firms from the INFOCAMERE and AIDA data, respectively. Columns I, III, V, VII, IX, XI, XIII and XV show information from the INFOCAMERE dataset, while all others columns refer to the AIDA dataset. INFOCAMERE data have been elaborated by Bologna Chambers of Commerce (Bureau of statistics).

5 Firms' entry and "involuntary" exit: our definition

We here proceed by explaining how we generated the variables related to the firms entry and exit, in particular, distinguishing between "voluntary" and "involuntary" exit from the market. Then, for the sake of completeness, we analyze the coverage of AIDA with respect to INFOCAMERE data in terms of entrants, exiting and active firms.

Once the dataset is in the standard long format (see Section 3), regarding the variables related to firms demography, we built firms entry indicator based on the incorporation year that we identified on the basis of the variable *date of incorporation*, which is provided by the AIDA dataset. For the investigation of firms' death as well as for industrial policy, it is very relevant to be able to distinguish the causes of firms exit. Ideally, the economist would like to have two distinct sets of motives for firms exit. One is related to the deliberate decision to cease the activity, relocate the business or successfully exit the market through acquisition. This is generally referred to as "voluntary" exit. The other

Table 7: INFOCAMERE and AIDA datasets: geographical distribution

Year	North-West		North-East		Center		South		Isles	
	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)	(VIII)	(IX)	(X)
2005	34.40	33.49	22.46	23.44	20.16	24.22	16.98	13.63	5.99	5.23
2006	33.91	33.30	22.34	23.28	20.38	24.37	17.28	13.80	6.10	5.24
2007	33.39	33.09	22.11	23.15	20.79	24.48	17.51	14.01	6.21	5.28
2008	31.26	32.89	20.34	22.98	25.61	24.51	16.86	14.28	5.93	5.35
2009	31.08	32.61	20.16	22.80	25.21	24.59	17.40	14.57	6.15	5.43
2010	30.92	32.48	20.08	22.78	25.06	24.61	17.62	14.67	6.32	5.47
2011	30.64	32.43	20.00	22.74	25.03	24.58	17.84	14.76	6.48	5.48
2012	30.30	32.41	19.87	22.65	25.11	24.54	18.08	14.87	6.64	5.53
2013	30.01	32.33	19.74	22.47	25.11	24.48	18.38	15.10	6.76	5.62
2014	29.70	32.91	19.54	22.09	25.13	24.58	18.72	15.60	6.90	5.81

Note. Columns I and II show the percentage of firms in the North-West of Italy in INFOCAMERE and AIDA data respectively; columns III and IV exhibit the share of firms operating in the North-East of Italy in INFOCAMERE and AIDA datasets, respectively; columns V and VI display the fraction of firms operating in the Center of Italy in INFOCAMERE and AIDA data, respectively; column VII and VIII show the share of firms operating in the South of Italy in INFOCAMERE and AIDA data, respectively; column IX and X exhibit the percentage of firms operating in the Isles in INFOCAMERE and AIDA datasets, respectively.

comprises all sort of events leading to firms exit against the willingness of the ownership. This second set of events leading to exit is commonly referred to as “involuntary”. In practice, one has to assign events leading to exit to one of the two categories, even if not always there is a sharp distinction between “voluntary” and “involuntary” exit. We focused on “involuntary” exit and based our definition of firms’ death on the type of administrative procedures a firm underwent. In particular, the variable *pending administrative procedures* identified 66 different administrative procedures; we resume in Table 8 the administrative procedures that we accounted for in order to define the “involuntary” firms’ exit indicator and in Table 9 all administrative procedures that we did not consider in our exit definition.

As Table 8 exhibits, we included the following administrative procedures that unambiguously lead to “involuntary” exit: bankruptcy, cancellation due to communication of allocation plan, cancellation ex officio from the register of companies, cancellation from the register of companies, composition with creditors, compulsory administrative liquidation, court order of cancellation, failure to meet prerequisites, impossibility of fulfillment of the company object, initial failure to meet the prerequisites for a company, no longer meets requirements specified for companies, post-bankruptcy composition with creditors, removal ex officio, supervening failure to meet the prerequisites for a company and winding up by official order. Therefore, we defined a firm as exited if it underwent one of the administrative procedures listed above. As a matter of fact, and as showed in Table 9 we did not include in the category of “involuntary” firms’ exit: “voluntary” exit (e.g. “approved by all partners” and “voluntary liquidation”), firms’ change of sector or province (e.g. “cessation of business within a province” and “transfer to another province”) and merger and acquisition (included “demerger”, “duplication”, “contribution”, “lease of company” and “transfer of firm”, among others). Moreover, we did not account for administrative procedures which do not unequivocally lead to “involuntary” exit. For example, we did not include “liquidation” and “closure due to bankruptcy or liquidation”, among others. Note that liquidation can be both voluntary and involun-

Table 8: List of administrative procedures which lead to “involuntary” firm exit

Bankruptcy
Cancellation due to communication of allocation plan
Cancellation ex officio of registration with register of companies
Cancellation from the register of companies
Cancelled ex officio pursuant to Article 2490 of the Italian Civil Code
Cancelled ex officio pursuant to Italian Presidential Decree no. 247 of 23 July 2004
Closure due to bankruptcy
Composition with creditors
Compulsory administrative liquidation
Conclusion of bankruptcy procedures
Court order of cancellation
Failure to meet prerequisites
Impossibility of fulfillment of the company object
Initial failure to meet the prerequisites for a company
Initiation of cancellation procedure
No longer meets requirements specified for companies
Post-bankruptcy composition with creditors
Removal ex officio
Removal ex officio, lack of tax code (Article 21 of Italian Presidential Decree no. 605 of 29 September 1973, as amended)
Removal ex officio following report by Provincial Handcraft Commission
Removal ex officio following report by register of companies for the registered office
State of insolvency
Supervening failure to meet the prerequisites for a company
Winding up by official order

tary; thus, we decided not to account for “liquidation” and “closure due to bankruptcy or liquidation” because the data did not provide any other specification and did not allow to make a distinction between voluntary and involuntary liquidation.

In order to correctly identify the time of exit from the market, we developed the following steps. We first looked at the year of the beginning of these administrative procedures.²⁹ Moreover, we complemented this information anticipating the year of exit to the last year in which the firm reported the balance sheet (this information was provided by the year identified by the variable *last accounting closing year*). We decided not to consider firms which did not report any relevant administrative procedure but that exited from the dataset at some point during the period of analysis.³⁰

We also had to deal with some problems related to the validity of information provided by *date of incorporation* and *last accounting closing year*. For instance, for some firms the year in which the firm underwent an administrative procedure resulted to be prior to their incorporation year. For these firms, we replaced the incorporation year with missing value, if the dataset provided balance sheet data even for the years before. If, instead, the dataset provided balance sheet data only for the following years, we considered the incorporation year as valid. In this latter case, we counted as valid administrative procedure the

²⁹This information was provided by the year identified by the variable *beginning of administrative procedure*. If a firm underwent more than one procedure generating an “involuntary” exit, we imputed the firm’s exit to the year associated to the first relevant administrative procedure.

³⁰It is worth noting that AIDA deletes the companies from the database if they do not report their balance sheet in the last 5 years. We did not account for these firms neither as active nor as exited; thus, for these firms, the “involuntary” exit variable assumes missing values.

Table 9: List of administrative procedures not included in the definition of “involuntary” exit

Annulment of entry
Annotation following communication by Provincial Handcraft Commission (Article 5 of Italian Law no. 443 of 8 August 1985)
Approved by all partners
Cancellation ex officio following creation of Chamber of Commerce, Industry, Craft Trade and Agriculture for Fermo
Cancellation ex officio following creation of Chamber of Commerce, Industry, Craft Trade and Agriculture for Monza
Cessation of any business
Cessation of business within the province
Closure due to bankruptcy or liquidation
Closure due to liquidation
Closure of local branch
Conclusion of liquidation
Contribution
Controlled administration
Court ordered administration
Court ordered liquidation
Court ordered seizure
Creation of new Chamber of Commerce, Industry, SME and Agriculture
Debt restructuring agreements
Demerger
Duplication
Extraordinary administration
Failure to re-establish multiple partners
Following expiry of time limits
Fulfilment of company object
Lease of company
Liquidation
Merger by incorporation into another company
Merger by incorporation of new company
Other reasons
Precautionary seizure of shares
Reason not specified
Removed ex officio because already included in the register of firms and not transferred to the register of companies
Transfer of firm
Transfer to another province
Transformation into a registered office
Transformation of legal status
Voluntary liquidation
Winding up
Winding up and liquidation
Winding up and placing into liquidation
Winding up in advance without liquidation
Winding up without liquidation

first procedure undergone after the firms' incorporation year. Moreover, we could not account for the incorporation year of 113 firms as it occurred after the registration of the last balance sheet (after these steps the number of firms without information on their incorporation year increased from 989 to 1120). Finally, due to the incompleteness of the data, we removed firms entered and exited in the same year.

After the steps described above, we obtained an unbalanced panel of 1,291,548 firms over the period 2004-2014. Out of this total, 923,205 firms had complete information on both entry and exit or were surviving at the end of the sample period (830,764 survived and 92,441 exited). Of the remaining cases, 423 firms did not have information on their entry (of these firms, 411 exited and 12 survived up to the end of the observed period), 367,357 firms did not have information on their exit (all these firms did not report any relevant administrative procedure and exited from the AIDA dataset at some point before 2014) and 563 firms did not have information on neither entry nor exit.³¹

Unfortunately, although financial data were available for the entire period 2005-2014 and firms' exit data concerned the period 2004-2014 in the AIDA database, valuable information on firms' exit only covered the years 2010-2013. Indeed, as suggested by the Bureau van Dijk division, the BvD section about pending administrative procedures, that we used in order to define "involuntary" firms' exit, has been added in AIDA only in December 2010. Thus, in order to perform empirical analysis on firms demographics, the researcher should focus on the post-crisis period (from 2010 onwards). Moreover, preliminary exploratory analysis revealed that, in most cases, there was a reporting lag of about two years; hence, in the AIDA dataset firms' exiting information were available for the period 2010-2013.³²

As we did for the size, sectoral and geographical distributions of firms, we now compare the coverage of AIDA with respect to INFOCAMERE data in terms of entrants, exiting and active firms. It is worth noticing that INFOCAMERE data do not distinguish between "voluntary" and "involuntary" exit. In particular, they consider a firm as exited if it underwent one of the administrative procedures listed in Tables 8 and 9. Accordingly, for an accurate comparison of the two datasets, we considered the same definition as in INFOCAMERE data for a firm exit.³³ Table 10 shows the comparison for the period 2009-2013. As already known from Table 5, Table 10 shows that the coverage of AIDA is lower than INFOCAMERE dataset (about 700,000 vs more than 900,000 firms).

The entry rate identified in the AIDA dataset is slightly higher than that reported by INFOCAMERE (9.381% vs 8.791% on average over the period 2009-2012), with the exception of the last year of investigation (2013), where entry rate are quite similar in the two sources of data. This one percentage point difference on average could be explained by the different number of active firms included in the two datasets. As for the exiting rate, for the period 2010-2012 the share of exits is higher in AIDA than in INFOCAMERE data, while in 2013 it is slightly lower. This difference could derive again from the different number of active firms and from the way in which the year of exit has been identified.

Despite the slight differences between the two datasets, AIDA provides a significant improvement for the analysis of firms demography with respect to many dataset previ-

³¹Even in this case information on firms' exit were not available because these firms did not report any relevant administrative procedure and exited from the AIDA dataset at some point before 2014.

³²As we mentioned before, we identified the same reporting lag even for financial variables included in the dataset.

³³The only difference between firms exit variable built for the comparison with INFOCAMERE data and "involuntary" exit indicator was the different set of administrative procedures which led to firms death.

Table 10: Comparison between AIDA and INFOCAMERE datasets: Firms' entry and exit

year	Number of active firms		% of entrants		% of exits	
	(AIDA)	(INFOC.)	(AIDA)	(INFOC.)	(AIDA)	(INFOC.)
2009	668458	903666	9.594	9.294	2.285	5.113
2010	673778	929340	10.315	9.504	6.681	5.288
2011	672522	953949	9.188	8.464	6.920	5.409
2012	671586	966141	8.427	7.901	6.512	5.657
2013	683940	982943	8.540	8.543	5.157	5.410

Notes. Columns II and III show the total number of active firms (limited liability companies) in AIDA and INFOCAMERE data, respectively; columns IV and V display the share of entrants. Columns VI and VII exhibit the share of exiting firms.

ously employed (i.e. Bottazzi et al., 2011; Cefis and Marsili, 2006; Esteve-Pérez et al., 2010; Varum and Rocha, 2012; Wagner, 2013, among others).

Granted all the caveats above, the procedures described in this section enabled to get a close track of the firm life-cycle, allowing to identify firm birth and, even more interestingly, making it possible to distinguish between “voluntary” and “involuntary” exit. This latter feature is not available in most firm-level dataset as it requires a tracking of events that is not always easy to carry out. This work provides, as far as we know, the first reference to achieve this indicator of firm exit which is of great importance both for the researcher as well as for the policy analyst.

6 Merging firm-level data with patents and trademarks

Firms are increasingly involved in the management of intellectual property rights, henceforth IPR, both in the forms of patents and trademarks. Furthermore, such activities are increasingly carried out also by firms traditionally classified as non-manufacturing. The activities that result in filing for the registration of a patent or a trademark enable to capture, although with some limitations, a relevant dimension of business dynamism that might well be related to firm growth, and more in general to firm demography.

In this respect, our aim is to obtain a dataset which allow to investigate how and the extent to which innovative activities might affect firms' entry and exit dynamics in addition to standard firms' performance variables. In order to do that, we employed patents and trademarks as proxies of innovations. More precisely, we linked the AIDA dataset described above with two separate datasets containing information on trademarks and patents owned by Italian firms, again provided by one of the sources of BvD. In this section we illustrate the procedure to merge the AIDA data with information on both patents and trademarks.

6.1 Patents data

Through BvD-AMADEUS it is possible to access information on patents hold by Italian firms,³⁴ and it turns out that there are more than 20,000 firms which own at least one patent (granted and non-granted) independently of where the patent has been applied to. More in general, AMADEUS provides some relevant information, including international patent classification (IPC) code, the application date, the number of citing documents and whether a patent has been granted or not, among others.

In order to obtain a suitable proxy of firms' innovative propensity, we only considered data on granted patents, with information on their application date, that have been applied at the United States Patent and Trademark Office (USPTO), at the European Patent Office (EPO), and/or at the Italian Patent and Trademark Office (IPTO). As a result, we restricted the sample of interest to 15,789 firms which own 97,540 patents. Among these, less than 5,000 firms only own patents applied at the national level.³⁵ Notice that the choice of including patents applied at the IPTO has been driven by the low propensity of Italian firms, in general, to apply for patents on international markets. The possibility to distinguish between patents applied internationally and patents applied only at the national level allowed us to account for their potential differences in terms of quality.

Notice that when collecting patent data the unit of analysis was the patent and not the firm. Further, a single patent entry might span over more than one single row as, for instance, there was more than one owner of the patents, or because the patent was relevant to more than one IPC (International Patent Classification) code. Data on patents, hence, needed to be re-arranged before they could be merged with standard firm-level data. Patents were uniquely identified by their application number,³⁶ while owner firms were identified by their BvD ID number. We identified patents applied at USPTO and/or EPO, and at IPTO by the first two letters of their publication number³⁷ and we dropped all other patents from the original dataset. Moreover, for each patent, we identified the application year by the four digits indicating the year of filing included in their application number.³⁸

Looking at the period 2004-2014 (the period covered by the AIDA dataset), we generated separate time-varying variables capturing the number of filed patents and the number of granted patents for each firm, distinguishing between patents applied at the USPTO and/or EPO (international level) and those applied at the IPTO (national level), respectively.³⁹ In particular, concerning the number of filed patents, we counted, for each year,

³⁴AMADEUS is a dataset of financial and business information for public and private companies across Europe. The dataset includes relevant information, such as annual balance sheet data, sectoral activities, patents and trademarks, among others.

³⁵Some patents are owned by more than one firm; in these cases we associated the patents to each owner, as suggested by the existing literature.

³⁶A patent could be applied to more than one office. Each office autonomously associate an application number and a publication number to the patent. There is not an indicator that reveals whether a patents has been registered to more than one office; thus, it is possible that we counted three times a patent if it was applied to EPO, USPTO and IPTO, respectively.

³⁷We only kept patents with publication number containing, as first two characters, "EP", "US" or "IT", indicating patents applied at the EPO, at the USPTO, or at the IPTO, respectively. Alternatively, we could have used the first two letters appearing in the patents application number.

³⁸The application number is made up of the country code, first two letters, the year of filling, four digit, and a serial number, which can assume a variable number of characters.

³⁹It has been possible by considering only rows with not missing BvD ID and by sorting data by BvD ID and application years; in this way we obtained for each firm one row for each patent owned.

the number of patents that each firm applied during the year of interest. We generated 22 variables identified by the variable label `FILED_PATENTS_USEP_` and `FILED_PATENTS_IT_` followed by the reference year (e.g. `FILED_PATENTS_USEP_2004`, `FILED_PATENTS_IT_2004`, ..., `FILED_PATENTS_USEP_2014` and `FILED_PATENTS_IT_2014`).⁴⁰ Similarly, we repeated the procedure to get the number of granted patents owned by firms in each relevant year (stock of granted patents), distinguishing between patents granted at the international and at the national level. In order to define the stock of granted patents for each firm, in each year, we did not account for patents applied more than 20 years before the year of interest (i.e. if a patent was applied in 1991 by a firm, we included this patent in the count of firm’s granted patents from 2004 to 2010 and we did not account for this patent in the firms’ stock of granted patents from 2011 onwards). This choice allowed not to account for patents which are too “old” to adequately represent a valuable proxy of firms’ technological capabilities (de Rassenfosse and Jaffe (2017), among others, have highlighted the importance to account for the decline in patents’ value during the life of patented inventions).

Another relevant dimension that has been possible to capture resorting to IPR data is the degree of coherence between the domain of knowledge, as represented by the IPC of the patents and the main activity of the firm, as proxied by the ATECO 2007 sector.

As mentioned above, a single patent could be associated with more than one IPC classes. In order to unequivocally assign each patent to a unique IPC code, we employed the following strategy: from the original IPC code (the complete IPC classification code comprises the combined symbols representing the section, class, subclass and main group or subgroup) we built the 4-digit IPC code (representing the section, class and subclass) and we kept as the unique one the 4-digit IPC code that recurred the most. Alternatively, when it was not possible to identify the most recurrent IPC code for a patent, we kept the first 4-digit IPC code that appeared in the original dataset from AMADEUS.

In order to built the concordance from IPC to economic sector, we relied on the probabilistic algorithm recently developed by Lybbert and Zolas (2014) which allows to build a correspondence between technological and production activities of firms at different levels. In particular, we linked IPC codes to ISIC codes (Rev. 4). Namely, we associated to each 4-digit IPC code the 4-digit ISIC code which displayed the highest probability weight.⁴¹

Focusing on the period 2004-2014, we built correspondence variables between patents’ technological field and firms’ economic sector. We extracted from the AIDA dataset the variables related to firms’ BvD ID number and their economic sector and we merged them with the AMADEUS dataset. Based on the 6-digit ATECO 2007 codes from the AIDA dataset, we generated for each firm the corresponding 4-digit, 3-digit and 2-digit ATECO 2007 codes. In order to verify whether there was equivalence between the ISIC code associated with the patents and the ATECO code associated with the firms, we used a correspondence table between 4-digit ISIC code and 3-digit ATECO 2007 code.⁴² Thus,

⁴⁰For multi-row firms that is, firms than own more than one patent, we repeated the values of these variables in each row.

⁴¹We merged the AMADEUS dataset with a table of concordance between 4-digit IPC codes and 4-digit ISIC codes resulting from the probabilistic algorithm provided by Lybbert and Zolas (2014).

⁴²We looked at the correspondence table between 4-digit ISIC code Rev 4 and 3-digit NACE Rev 2 available on line on the EUROSTAT website. Indeed, the ATECO 2007 classification is based on NACE Rev 2 classification and the two classifications are identical for the first 4 digits. Moreover, based on the correspondence table between 4-digit ISIC code Rev 4 and 3-digit NACE Rev 2, the 4-digit ISIC code 2100, 2410, 4100, 5510, 6810 7490 and 8510 correspond to more than one 3-digit NACE Rev 2 but

we generated, for each firm and each year, the correspondence variables between 4-digit ISIC code and 3-digit ATECO code, separately for patents applied at the USPTO or EPO and for patents applied at the IPTO. These variables assumed value one if at least one of the applied patents in the referred year showed an associated 4-digit ISIC code which was equivalent to the firm owners' 3-digit ATECO 2007 code (we generated 22 variables identified by the variable label `CORR_FILED_USEP3_` and `CORR_FILED_IT3_` followed by the referred year and we repeated the values of these variables in all rows for multi-row firms). Following the same approach, we created the correspondence variables between 4-digit ISIC code and 3-digit ATECO code considering the stock of granted patents owned by firms in each year (we generated 22 variables identified by the variable label `CORR_VALID_USEP3_` and `CORR_VALID_IT3_` followed by the referred year and we repeated the values of these variables in all rows for multi-row firms). Finally, given that in the AIDA dataset some firms (24,358 firms) only had 2-digit ATECO code, following the same approach as described above, we generated the correspondence variables for applied and granted patents, in each year for each firm, considering the equivalence between 4-digit ISIC code associated to each patent and 2-digit ATECO code associated to each owner (we created 44 variables identified by the variable label `CORR_FILED_USEP2_`, `CORR_FILED_IT2_`, `CORR_VALID_USEP2_` and `CORR_VALID_IT2_` followed by the referred year).

After the creation of these relevant variables, we kept only one row for each firm and then proceeded with the reshape command to convert the data from wide to long format (in the long format years of data, for each firm, is displayed in separate rows). The reshaped file contains information on the number of filed and granted patents, respectively at the national and international level, on the correspondence between patents technological fields and firms' economic sectors for 15,789 firms over the period 2004-2014 (out of these 15,789 firms only 15,137 are even included in the AIDA dataset. The remaining 1,276,411 firms in the AIDA dataset do not own any patent).

6.2 Trademarks data

With regard to trademarks as proxy for firms' innovative activity, we got information from AMADEUS, which includes more than 20,000 Italian firms which own at least one trademark. For filed trademarks, AMADEUS provides some relevant information, including NICE classification code, the filing date and information on their registration, among others. In particular, we focused on registered trademarks that have been filed at the United States Patent and Trademark Office (USPTO) or at the Office for Harmonization in the Internal Market (OHIM).⁴³ Thus, we restricted our attention to 19,168 firms which own 59,431 trademarks.⁴⁴

to only one 2-digit NACE Rev 2 code, respectively. In AMADEUS dataset there were 9,138 patents, owned by 1,706 firms, that showed ISIC code equal to 2100, 2410 or 4100, while the other problematic 4-digit ISIC codes did not appear in the dataset. For the 4-digit ISIC code 2100, 2410, 4100 we built the correspondence between ISIC and ATECO considering 2-digit ATECO 2007 codes 21, 24 and 41, respectively.

⁴³Unfortunately, differently from patents, for trademarks AMADEUS does not allow to get information even on registered trademarks that have been applied at the national level.

⁴⁴As patents, even trademarks could be owned by more than one firm; in these cases we associated the trademarks to each owner, as suggested by the existing literature. Similarly to patents, trademarks could be filed to both USPTO and OHIM. Each office autonomously associate an identification number to the trademarks and there is not an indicator that reveals whether a trademark is registered to more than one office; thus, it is possible that we double counted a trademark if it was applied to both offices considered.

Importing firms’ trademarks data from AMADEUS to STATA we obtained more than one row for each firm, each row referring to a single trademark owned. Similarly to what we did for patents, we generated the number of filed and registered trademarks by firms in each year (2004-2014). In particular, concerning the number of filed trademarks in each year, we counted the number of trademarks that each firm applied during the year of interest. We generated variables identified by the variable label `FILING_TRADEMARKS_` followed by the referred year (e.g. `FILING_TRADEMARKS_2004`, `FILING_TRADEMARKS_2005`, ..., `FILING_TRADEMARKS_2014`).⁴⁵ Similarly, we generated indicators reporting the number of registered trademarks owned by firms in each year (e.g. `VALID_TRADEMARKS_2004`, `VALID_TRADEMARKS_2005`, ..., `VALID_TRADEMARKS_2014`). In order to define the stock of registered trademarks for each firm, in each year, we considered only trademarks applied before or in the year of interest, that expired after the referred year.⁴⁶

As for patents, even for trademarks one of our main objective is to obtain a final dataset which allows to develop a correspondence measure between trademarks technological fields and production activities of firms. We needed to keep information on NICE classification from the AMADEUS data.⁴⁷ Differently from patents, for each trademark the AMADEUS dataset provided only a single NICE class. In order to build the NICE-economic sector concordance we relied on the probabilistic algorithm recently developed by Lybbert et al. (2013) and we linked trademarks’ NICE codes to 2-digit ISIC codes (Rev. 4).⁴⁸ We extracted from the AIDA dataset the variables related to firms’ BvD ID number and their economic sector and we merged them with the AMADEUS dataset. Based on the 6-digit ATECO 2007 codes from the AIDA dataset, we generated for each firm the corresponding 4-digit, 3-digit and 2-digit ATECO 2007 codes. Hence, we generated, for each firm and each year, the correspondence variables between 2-digit ISIC code associated with each trademark and firms owners’ 2-digit ATECO code. These variables took value one if at least one of the applied trademarks in the referred year showed a 2-digit ISIC code which was equivalent to the firm owner 2-digit ATECO 2007⁴⁹ (we generated variables identified by the variable label `CORR_FILED_TRADE_` followed by the referred year and we repeated the values of these variables in all rows for multi-row firms). Following the same approach, we created the correspondence variables between 2-digit ISIC code and 2-digit ATECO code considering the stock of registered trademarks owned by firms in each year (we generated variables identified by the variable label `CORR_VALID_TRADE_` followed by the referred year and we repeated the values of these variables in all rows for multi-row firms).

After generating these “correspondence” variables, we kept only one row for each firm and we proceeded with the reshape procedure converting the data from wide to long format (in the long format years of data, for each firm, is displayed in separate rows). The reshaped file contains information on the number of filed and registered

⁴⁵For multi-row firms, thus for firms that own more than one trademark, we repeated the values of these variables in each row.

⁴⁶The trademarks expiration date was only available for trademarks applied at the OHIM, thus as suggested by the BvD division, we considered all registered trademarks applied at the USPTO as valid up to the beginning of 2015, the time of the last available update for USPTO data.

⁴⁷The NICE classification is a 2-digit international classification of goods (codes from 1 to 34) and services (codes from 35 to 45) applied for the registration of trademarks that has been adopted by the Nice Agreement (1957).

⁴⁸The probabilistic algorithm proposed by Lybbert et al. (2013) does not allow to go further than using a 2-digit code for firms’ economic sectors.

⁴⁹ISIC Rev 2 and ATECO 2007 classifications are identical for the first 2 digits.

trademarks, on the correspondence between trademarks technological fields and firms' economic sectors for 19,168 firms over the period 2004-2014 (of these 19,168 firms only 19,141 are even included in the AIDA dataset). Finally, we merged the AIDA dataset with both AMADEUS datasets containing information on firms' patents and trademarks, respectively. In the final dataset, which includes 1,291,548 firms over the period 2004-2014, 3,573 firms own both patents and trademarks, 11,564 firms only own patents, 15,568 firms only own trademarks and 1,260,843 firms do not own neither patents nor trademarks.

7 Conclusion

In this paper, we have described the most recurrent issues related to the building of a firm-level dataset and how they can be addressed. The set of procedures that we propose here is applied to the specific case of Italian limited liability companies as tracked by Bureau van Dijk (BvD) AIDA, but the methods that we suggest here is however far more general. In this work we also suggest how to employ data on events affecting the operation of business firms to infer relevant information on business demography. As far as AIDA is concerned, it is possible to infer information on firm entry and age by resorting to the year in which the firm first appeared in the business register. In a similar manner, it is also possible to identify the exit of the firm and, far more relevant, also to distinguish between "voluntary" and "involuntary" exit. Moreover, merging the AIDA dataset with information on firms' granted patents and registered trademarks - in our case provided by BvD AMADEUS - allows to further investigate the determinants of firms' likelihood to survive, focusing on the role played by the innovation activities.

In a more general perspective, this paper introduces a series of procedures that the researcher might apply in order to build firm-level dataset starting from different sources of data collected by National Statistical Offices or other institutions. Finally, by proposing a homogeneous set of procedures, we believe that our work also contributes to the replication of empirical analyses performed on the same set of data by different researchers.

References

- BAILY, M. N., C. HULTEN, D. CAMPBELL, T. BRESNAHAN, AND R. E. CAVES (1992): "Productivity dynamics in manufacturing plants," *Brookings papers on economic activity. Microeconomics*, 1992, 187–267.
- BALDWIN, J. R. AND M. RAFIQUZZAMAN (1995): "Selection versus evolutionary adaptation: Learning and post-entry performance," *International Journal of Industrial Organization*, 13, 501–522.
- BARTELSMAN, E. J. AND M. DOMS (2000): "Understanding productivity: Lessons from longitudinal microdata," *Journal of Economic Literature*, 38, 569–594.

- BOTTAZZI, G., M. GRAZZI, A. SECCHI, AND F. TAMAGNI (2011): “Financial and economic determinants of firm default,” *Journal of Evolutionary Economics*, 21, 373–406.
- CEFIS, E. AND O. MARSILI (2006): “Survivor: The role of innovation in firms survival,” *Research Policy*, 35, 626–641.
- DAVIS, S. J., J. HALTIWANGER, AND S. SCHUH (1996): “Small business and job creation: Dissecting the myth and reassessing the facts,” *Small business economics*, 8, 297–315.
- DE RASSENFOSSE, G. AND A. B. JAFFE (2017): “Econometric Evidence on the R&D Depreciation Rate,” NBER Working Papers 23072, National Bureau of Economic Research, Inc.
- DISNEY, R., J. HASKEL, AND Y. HEDEN (2003): “Entry, exit and establishment survival in UK manufacturing,” *The Journal of Industrial Economics*, 51, 91–112.
- DOSI, G., M. GRAZZI, L. MARENGO, AND S. SETTEPANELLA (2016): “Production theory: accounting for firm heterogeneity and technical change,” *The Journal of Industrial Economics*, 64, 875–907.
- ESTEVE-PÉREZ, S., A. SANCHIS-LLOPIS, AND J. A. SANCHIS-LLOPIS (2010): “A competing risks analysis of firms exit,” *Empirical Economics*, 38, 281–304.
- GRAZZI, M., R. SANZO, A. SECCHI, AND A. ZELI (2013): “The building process of a new integrated system of business micro-data 1989–2004,” *Journal of Economic and Social Measurement*, 38, 291–324.
- GRILICHES, Z. AND J. MAIRESSE (1999): “Production Functions: The Search for Identification,” in *Econometrics and Economic Theory in the Twentieth Century: the Ragnar Frisch Centennial Symposium*, ed. by S. Steiner, Cambridge University Press: Cambridge.
- ISTAT (2015): “Struttura e competitività del sistema delle imprese industriali e dei servizi,” Statistiche report, ISTAT.
- KALEMLI-OZCAN, S., B. SORENSEN, C. VILLEGAS-SANCHEZ, V. VOLOSOVYCH, AND S. YESILTAS (2015): “How to construct nationally representative firm level data from the ORBIS global database,” NBER Working Papers 21558, National Bureau of Economic Research, Inc.
- LYBBERT, T. J. AND N. J. ZOLAS (2014): “Getting patents and economic data to speak to each other: An “Algorithmic Links with Probabilities” approach for joint analyses of patenting and economic activity,” *Research Policy*, 43, 530–542.
- LYBBERT, T. J., N. J. ZOLAS, AND P. BHATTACHARYYA (2013): “An “Algorithmic Links with Probabilities” Concordance for Trademarks: For Disaggregated Analysis of Trademark and Economic Data,” WIPO Working Papers 14, World Intellectual Property Organization-Economics and Statistics Division.
- SYVERSON, C. (2011): “What determines productivity?” *Journal of Economic Literature*, 49, 326–365.

VARUM, C. A. AND V. C. ROCHA (2012): “The effect of crises on firm exit and the moderating effect of firm size,” *Economics Letters*, 114, 94–97.

WAGNER, J. (2013): “Exports, imports and firm survival: First evidence for manufacturing enterprises in Germany,” *Review of World Economics*, 149, 113–130.



Alma Mater Studiorum - Università di Bologna
DEPARTMENT OF ECONOMICS

Strada Maggiore 45
40125 Bologna - Italy
Tel. +39 051 2092604
Fax +39 051 2092664
<http://www.dse.unibo.it>