BANCA D'ITALIA

EUROSISTEMA

Questioni di Economia e Finanza

(Occasional Papers)

Matching of PATSTAT applications to AIDA firms: discussion of the methodology and results

by Francesca Lotti and Giovanni Marin

June 2013

Number 166

BANCA D'ITALIA
EUROSISTEMA

Questioni di Economia e Finanza

(Occasional papers)

Matching of PATSTAT applications to AIDA firms: discussion of the methodology and results

by Francesca Lotti and Giovanni Marin

Number 166 – June 2013

The series Occasional Papers *presents studies and documents on issues pertaining to the institutional tasks of the Bank of Italy and the Eurosystem. The* Occasional Papers *appear alongside the* Working Papers *series which are specifically aimed at providing original contributions to economic research.*

The Occasional Papers *include studies conducted within the Bank of Italy, sometimes in cooperation with the Eurosystem or other institutions. The views expressed in the studies are those of the authors and do not involve the responsibility of the institutions to which they belong.*

The series is available online at www.bancaditalia.it .

ISSN 1972-6627 (print)
ISSN 1972-6643 (online)

Printed by the Printing and Publishing Division of the Bank of Italy

MATCHING OF PATSTAT APPLICATIONS TO AIDA FIRMS: DISCUSSION OF THE METHODOLOGY AND RESULTS

by Francesca Lotti* and Giovanni Marin**

Abstract

This paper is a brief methodological note on the matching of Italian firms in the AIDA database with applicants at the European Patent Office from the PATSTAT database. The need to match data on patent applications with balance-sheet information stems from the importance of patent statistics as a source of information on the innovative performance of firms. Starting from recent efforts to match applicants in PATSTAT with firms in the Bureau van Dijk databases (ORBIS, AMADEUS, FAME), we added an improved cleaning routine to maximize exact matches, followed by an approximate matching based on multiple combination of similarity scores. Starting with 272,475 firms, we matched 49,369 EPO applications in the period 1977-2009. The matching covers 68 percent of EPO applications by Italian firms for the entire period and 89 percent for 2000-2009. Finally, we describe the time, sector, size, geographical location and technology distribution of the matched applications.

JEL Classification: C81, O31, O34.
Keywords: names harmonization, patents, approximate matching, PATSTAT, AIDA.

Contents

* Banca d'Italia, Via Nazionale, 91, Roma, francesca.lotti@bancaditalia.it

** CERIS-CNR, Via Bassini, 15, Milano, Italy, e-mail: g.marin@ceris.cnr.it

1 Introduction

The use of patent data as a measure of innovation output, as opposed to input measures like R&D expenditure, has long been proposed by economists (Comanor and Scherer, 1969). Patent data have manifold advantages and also limitations[1]. The applied economic literature on innovation patterns has used a number of gauges: significant inventions (Pavitt (1984) investigates the sectoral patterns of technical change examining some 2,000 significant innovations in Britain since 1945); share of innovative products; sales at the firm level (Crepon et al., 1998; Griffith et al., 2006); and binary measures of any process or product innovation at micro level (Griffith et al., 2006)[2]. Unlike these measures, patent data are collected for the entire set of patents and they are objective, with no room for bias due to self-reported measures.

An issue in using patent data is their integration with other micro data, which is complicated by internal problems and patent database inconsistencies. Patent data are collected for various legal and administrative purpose, with no specific methodological requirements. This lack of standardization poses a number of problems for statistical analysis: (i) lack of a unique identifier for applicants and inventors; (ii) typing mistakes in textual fields such as name or location and (iii) numerous observations with missing information (application or publication date, applicant's or inventor's name or location, IPC class). Given the huge numbers of patents, inventors and firms in the databases, these problems increase the cost[3] of using patent data at the micro level.[4]

A first systematic attempt to integrate patent data at firm level with other micro is the National Bureau of Economic Research's productivity program from 1978 through 1988 (Bound et al., 1984; Hall et al., 1988). This NBER project was designated to build an integrated database on listed U.S. manufacturing firms with data on balance-sheet, income statement, R&D and patent for the study of innovation patterns, productivity and firm value at micro level. Starting with a panel of about 2,600 large manufacturing firms available in Compustat, researcher matched about 300,000 patent applications to the U.S. Patent and Trademark Office in the period 1965-1981[5]. They combined name harmonization with visual matching, in order to minimizing both false positives and false negatives. After a first round based on exact matching and rough approximate matching, Hall et al. (1988) visually checked all matches and performed recursive visual matching of all possible unmatched firms and applicants. They repeated this procedure for each update of the Compustat and USPTO data from 1978 through 1988. This procedure, though effective, is quite costly (in money and time) and hard to extend to databases including small and medium-sized enterprises. The matching between USPTO applicants and companies in Compustat has been updated (Hall et al., 2001; Cockburn et al., 2009).

The preliminary and final datasets produced by this project have been used in very influential articles published by researchers affiliated to the NBER. Pakes and Griliches (1980) investigate the relationship between R&D expenditures and patent counts for 121 large corporation in 1968-1975. They find a very strong cross-sectional correlation and a significant, though weaker, correlation within firm. They also investigate the extent to which past R&D affects current patent counts, finding strong contemporaneous correlation and weaker (though still significant) positive correlation with past R&D expenditures. Hausman et al. (1984) develop an econometric method aimed at investigating patent count data in a panel setting. They apply their method to the relationship between R&D expenditures and patent counts at the firm level[6] to determine the lag with which R&D affects innovation as measured by patents. Hall et al. (1986) extend the analysis to a larger panel (642 firms) but a shorter period (1972-1979). Finally, Griliches et al. (1988) go beyond the simple R&D-patent relationship to inquire into the relationship between innovation, stock market value, and the value of patents exploiting data on patent renewal fees and the assessment of knowledge spillovers.

[1] For a more detailed review of the literature the relevance of patent statistics see Pavitt (1985), Griliches (1990) and OECD (2009).

[2] The use of innovative sales and binary measures has been favoured by the inclusion of these measures in the questionnaire of the various waves of the Community Innovation Survey (CIS). On the use of CIS data in microeconometric analysis of innovation patterns, refer to Mairesse and Mohnen (2010).

[3] Standardization, disambiguation and matching of patent data generate a series of costs for the researcher: apart from the monetary cost for the staff, time lags between the start of the project and the moment when statistical analysis becomes possible, as well as inaccuracies in standardization, disambiguation and matching, hence measurement errors and biases.

[4] At the macro, regional and sectoral level, systematic data are available from OECD.

[5] At the same time, the USPTO reported patents granted only.

[6] The sample used by Hausman et al. (1984) included in 128 firms for the period 1968-1974.

Lotti and Schivardi (2005) matched European Patent Office (EPO) patent applications to firms in a small version of the AMADEUS database (about 115,000 firms) for the EU15, taking exact matches and visually-checked matches via the SOUNDEX algorithm.[7] The authors acknowledge that they fail to match many EPO applications, thus producing a good share of 'false zeros'. Moreover, they neglect the possibility of different firms having the same name (either in AMADEUS or in the patent database). Finally, they do not exploit data on the location of firms and applicants to improve precision.

More recent efforts to integrate patent data with other firm-level data are (Thoma and Torrisi, 2007; Thoma et al., 2010). The idea was to devise automatic routines and algorithms to harmonize, disambiguate and match applicants with firms. They matched patent applications at USPTO and EPO available in the Worldwide Patent Database (PATSTAT), with companies in the AMADEUS database (Bureau van Dijk)[8]. Unlike the NBER project, the approach taken here allows for small but significant share of false matches and false negatives; however, it has the advantage of covering SMEs[9] and using more efficient automated methods. A major effort was made to create routines for name harmonization to correct the most common typing mistakes and standardize name conventions [10]. This produced good results from exact matching made possible effective approximate matching possible. In approximate matching, information on the location of applicants and firms is used to screen out false matches, and scores are computed both in terms of string similarity functions and in terms of token distance (see Thoma et al. (2010) for further details). Due to the huge number of applicant - firm matches, no global visual check is performed.[11] Ultimately 131,065 companies included in AMADEUS were identified as EPO applicants corresponding to about a million of EPO applications in 1979-2008.

Helmers et al. (2011) study United Kingdom only, harmonize and match firms in FAME (*Financial Analysis Made Easy*, the British counterpart of AMADEUS, which includes all British registered firms in 2000-2007) with patent applications to EPO and IPO (Intellectual Property Office, the British patent office) and with trademarks. They consider only exact matches, with a great deal of effort going to names harmonization. For 2003, 83 percent of EPO applications were matched, 57 percent of IPO applications and 86 percent of trademarks by British business entities. The authors also give some descriptive analysis on the distribution of patent applications and trademarks by firm size, location, sector and technology. In section 3.2 we replicate part of this analysis for Italian firms in AIDA.

To conclude this brief review of the literature, we briefly describe the APE-INV project (Academic Patenting in Europe) aimed at matching inventors in the PATSTAT database with academic researchers and professors (Lissoni et al., 2010). The project, which has begun in June 2009, is headed by the KITeS (Centre for Knowledge, Internationalization and Technology Studies of the Bocconi University in Milan). It is funded by the European Science Foundation (ESF). Compared with to firm-applicant matching, inventor-researcher matching has more problems of disambiguation (high frequency of name-surname pairs in both lists) and fewer problems of name harmonization.

This paper describes the methodology and the results of the matching of Italian patent applicants (and applications) with Italian firms in the AIDA database (Bureau van Dijk). The combination of patent data with other non-survey data is likely to attenuate the high risk of selection bias that marks innovation surveys(Mairesse and Mohnen, 2010) although it limits the variety of research issues that might be addressed by comparison with innovation surveys. However, the AIDA database does not contain the population of Italian firms and there is a bias 'by construction' due to the exclusion of inactive firms after four years.

The rest of the paper is structured as follows. Section 2 discusses the main problems in matching PATSTAT with other databases of firms (2.1) and describes the data sources (2.2) and the methodology used for the matching (2.3). Section 3 discusses the results of the matching (3.1) and sets out some stylized facts that follow (3.2). Section 4 concludes.

[7] The SOUNDEX algorithm 'produces matches for strings using a weighting scheme, according to which each component of the string is assigned a certain weight and matches are produced accordingly' (Lotti and Schivardi, 2005).

[8] The version of AMADEUS employed by Thoma et al. (2010) covers 10 million companies for the years 1998-2006.

[9] The under-representation of SMEs in AMADEUS, although still significant, is considerably less marked than in COMPUSTAT.

[10] In order to increase the likelihood of identifying all matches, name variations of single entities applying for PCT/WIPO were included as an additional dictionary.

[11] Thoma et al. (2010) perform a visual check of approximate matches for a small sample of 76 applicant - firm pairs, finding just 3 false matches.

2 Data and methodology

2.1 Where is the missing link?

Matching data from different sources is a common problem in applied economics and researchers are required to devote considerable time and effort in consuming tasks not directly related to the research itself. Moreover, matches are frequently uncertain due to the lack of a unique identifier in different sources. This inconsistency can produce severe measurement errors, missing values (generally non-random), and small samples reducing the reliability of any estimate.

Data on patent applications are outside the scope of databases on firms, which generally cover balance-sheet information and 'demographic' data (year of incorporation, legal status, location, sector of activity, etc.).[12] Patent data are made available in specialized databases such as the 'EPO Worldwide Patent Statistical Database' (PATSTAT), released twice a year by the European Patent Office. But these databases have no unique identifier for applicants,[13] because their primary unit of analysis is the patent application. This causes problems in taking applicants as unit of reference even within patent databases, owing to:

- variations in a firm's name (due to actual name changes, change in name conventions, typing mistakes, merger and acquisitions);
- duplication of a name for different firms.

The data collected by patent offices are sometimes inconsistent. The names of applicants and inventors are often collected under different name conventions, without considering whether he was already reported in previous applications. And this missing temporal link makes typing mistakes more common. Finally, the absence of a unique identifier for applicants and inventors, together with the limited availability and lack of standardization of other information, such as addresses, makes it harder to distinguish between a duplication of names due to multiple applications by the same person or firm and the existence of distinct persons or firms with the same name.

The consequence is biases in statistics at the applicant/inventor level (applications count, citations count). Corporate applicants, which account for about 81 percent of Italian EPO applications between 1977 and 2009, are more prone to the problem of name variations while inventors are more prone to duplication. This gives rise to an expected negative bias (underestimation of applications/citations count), for corporate applicants and an expected positive bias for statistics at the inventor level, due to the high frequency of coincidence of common names.

When the names of applicants have to be harmonized and matched with external lists of names of firms, these problems are exacerbated, especially when these lists do not represent the entire population of firms. Harmonization may transform distinct names into identical harmonized names. When one of these false duplicates is not included in the list of names because the list itself represents just a fraction of the whole population, the harmonized applicant name may be matched to the wrong duplicate in AIDA. To minimize this risk, false matches may be detected by referring to additional information, such as location of both applicants and firms in AIDA.

2.2 Data

We use three sources of data: the AIDA database, the Worldwide Patent Database (PATSTAT) and the results of the matching by Thoma et al. (2010).

2.2.1 AIDA

AIDA is a commercial database on Italian firms, maintained by Bureau van Dijk. It gives balance-sheet, income statement and other information, such as location, sector, year of incorporation, ownership and equity participations in other firms, covering a 10-year time window. We use the AIDA top version.

[12] Longitudinal databases on companies are generally characterized by problems when recording changes in the status of the company. It is generally hard to have information the reasons behind the disappearance of a company from the database, either due to exit from the market, temporary inactivity, merger, acquisition, transformation or changes in the sampling strategy. However, in order to limit the issues at stake, we do not consider this issue in the current paper.

[13] Possible unique identifiers for firms are the Chamber of Commerce registration number and the tax code.

Unlike the full version (AIDA SMALL + MEDIUM + TOP) it covers only a small proportion (about 10 percent) of firms with turnover below 1.5 million euros, as the name implies it covers all those with turnover above that amount. AIDA TOP involves three types of selection. First, there is the selection of the full AIDA database of a million of the over 4 million firms active according to Istat. This selection is not explicitly disclosed by Bureau van Dijk. Second, within the full AIDA database, AIDA TOP severely screens out most small firms. Finally, firms that have been inactive for more than four years are generally deleted, substantially reducing the coverage in the first years of the database and introducing an additional selection bias, in that surviving firms are likely to differ from those that exited the market.

A final consideration is group/subsidiary status. We did not consolidate patent applications and financial accounts of subsidiaries in their corporate group.,

In April 2011, date in which we extracted the data, AIDA TOP includes 272,475 companies. Tables 1 and Table 2 give their absolute and relative sectoral distribution[14] by year. Values refer to firm/year pairs where no balance-sheet information is missing. The coverage (share of firms with non-missing balance-sheet information) increases from 86,142 firms (35.5 percent) in 2000 to 235,492 (88.8 percent) in 2007, owing to the entry of new firms, the extension of the coverage of existing firms, and the fact that newly inactive firms are generally not reported. The dynamics at the sector level is quite smooth except for *Finance, real estate*, where we find a significant increase in coverage starting in 2004 (the sector's share jumped from 3.8 percent in 2003 to 6.3 percent in 2004, with an increase of about 50 percent in absolute terms). This probably reflects the change in the selection rules for AIDA TOP in 2004. Figure 2 shows the geographical distribution (by province) of the firm-year pairs where no balance-sheet information is missing. Apart from the province of the largest cities (Milan, Rome, Turin, Naples, accounting for about 27 percet of all the pairs), high concentrations of firms are found in other provinces in Lombardia (Monza e Brianza, Varese, Bergamo, Como, Lecco), in Prato in Tuscany (especially in the textile sector); in the NorthEast (Trieste, Padova, Treviso, Vicenza, Verona, Venice) and in Emilia-Romagna around Bologna (Bologna itself, Rimini, Modena). Low density characterized most central and southern provinces.

2.2.2 PATSTAT

The EPO Worldwide Patent Statistical Database (dubbed PATSTAT) is prepared by the European Patent Office on behalf of the OECD Taskforce on Patent Statistics. It covers patent applications in more than 80 countries. The data include: (i) applicants' and inventors' names and addresses; (ii) title and abstract of patent applications; (iii) priority, patent families and PCT links; (iv) bibliographical information (citation links); (v) classification of patents by technology class.

We retrieved data from the April 2011 Release (PATSTAT is released in April and October every year) on all EPO applications[15] filed from 1977 through 2009. For each application (`appln_id`) we retrieved the application date and priority date[16] (`appln_date` and `prio_date`), applicant's name (`person_name` and `doc_std_name`) and address (`person_address`) and IPC class (`ipc_class_symbol`).[17] Total applications by Italians (results not shown but available upon request) follow a smoothly increasing trend until 2006, followed by a decline in 2007 and a sharp huge drop in 2008 (about to half of 2006) due to the well-known truncation problem (Hall et al., 2001). The truncation for the years close to the date of collection is due to delays in the publication of EPO applications. These can be published eighteen months after the application or priority date, leading to an underestimation of applications counts in the last three years.

[14] Macro-sectors are defined as follows (Nace Rev. 1.1 codes): Agriculture and Mining 01-14; Medium-High Technology Manufacturing 23-35; Low Technology Manufacturing 15-22 and 36-37; EGW (Electricity, Gas and Water supply), Construction 40-45; Wholesale, Retail, Hotels 50-55; Transport and Telecommunication 60-64; Finance, Real Estate 65-71; Computer 72; R&D services 73; Business activities 74; Other Services 75-95.

[15] De Rassenfosse and Wastyn (2012) note the possible bias stemming from the use of data for only (or a few) patent offices. We accordingly plan to extend the matching to patent applications to the Italian patent office and filed under the PCT treaty.

[16] The priority date is the date on which an application for a specific invention is filed. After this first application, the applicant can apply to other patent offices for the same invention within 12 months claiming protection for that invention since the priority date.

[17] We did not extract and use information on continuations and technical relation, so the raw counts of patent applications in the following section entail double counting of patented innovations of applications due to multiplicity of applicants for the same application and distinct applications (continuations, technical relations) for the same innovation.

2.2.3 Matching by Thoma et al. (2010)

The results of the matching done by Thoma et al. (2010) have been recently disclosed.[18] The authors matched 4,796 Italian firms in AMADEUS as applicants to the EPO corresponding to about 24,000 EPO applications (reference period: 1978-2006) and published the complete lists of harmonized names and locations of companies in AMADEUS and applicants in PATSTAT. This database has been used here tor two ends: (i) including the matches identified by the study; (ii) using the harnomized names together with their Bureau van Dijk identifier as additional name variations for our list of firms in AIDA.

2.3 Methodology

Much effort went into improving exact matching of EPO applicants in PATSTAT with firms in AIDA, with recursive rounds of harmonization and improvement of the cleaning routines. In addition, we extended coverage by including approximate matches, which were checked visually.[19]

The matching was performed in nine steps:

1. preliminary check, using a restricted sample, of the main problems of name harmonization both for the applicants in PATSTAT and the firms in AIDA;
2. recursive harmonization of names and improvement of the routines at each step;
3. identification of duplicates in the list of firms in AIDA;
4. exact matching of non-duplicate harmonized names;
5. harmonization of addresses;
6. exact matching of duplicate harmonized names using harmonized addresses;
7. identification of candidate pairs for the approximate matching and creation of similarity measures;
8. visual check of approximate matches;
9. inclusion of EPO applications matched by Thoma et al. (2010) and treatment of applications matched with multiple applicants.

The point of departure is the set of harmonization routines published in the new homepage of the NBER Patent Data Project,[20] which consists in a first set of general cleaning and standardization commands: elimination of punctuation, standardization of special characters, elimination of double spaces, transformation of lower cases into upper cases and unification of acronyms. A second set of commands sandardizes common name conventions. The standardization concerns both the legal status of the firm (for instance, SOCIETÀ PER AZIONI, SOC PER AZIONI, SOC PER AZ are all standardized as SPA) and other common words in company names that could be written or abbreviated in various ways (e.g. INDUSTRIES vs IND, MANUFACTURING vs MANUF vs MFG, INTERNATIONAL vs INT). Finally, a new string variable, the "stem name", is created by removing legal status and some common words and abbreviations (e.g. MFG, INT, IND). The stem name is then used by a Perl programme to identify candidate approximate matches.

These routines have been further adapted and improved to fit Italian names. We ran the same name harmonization routines on the list of applicants in PATSTAT[21] and on the list of firms in AIDA,[22] followed by routines to standardize addresses in AIDA and PATSTAT. Unlike AIDA, PATSTAT writes the address (street, number, postal code, city, country) in a unique field.[23] We did some basic cleaning

[18] http://www.researchoninnovation.org/epodata/

[19] The visual check, based on both harmonized and original names and addresses, serves to minimize errors in approximate matching. Candidate pairs with identical scores could be visually disambiguated even in absence of any measurable or standard automatic rule.

[20] https://sites.google.com/site/patentdataproject/

[21] We used both the list of applicants in the table `tls206_person` (field `person_name`) and the list in table `tls208_doc_std_nms` (field `doc_std_name`) of the PATSTAT database.

[22] We included any available past denominations, and also added all firm names already matched by Thoma et al. (2010).

[23] The postal code has been extracted by identifying, within the unique field of the address, all 5-digit numbers (without spaces).

on the addresses, focusing on common abbreviations (e.g. 'S.' instead of 'San' or 'Santo') and the English versions of major city names (Rome-Roma, Milan-Milano, Venice-Venezia, etc.). Finally, we identified all unique firms whose name recurred more than once in the list of AIDA firms (defined as duplicates).[24]

Once names and addresses were harmonized, we identified all exact matches with the same address in AIDA and PATSTAT. The criterion was matching either of the municipality (matching of the municipality reported in AIDA with any substring including the municipality name in the address in PATSTAT) or the postal code (or its reduced version with 3 or 4 digits). Within these matches, we visually checked duplicate names that shared a set of patent applications. These matches depend on coincidence of both name and location for firms in AIDA and applicants in PATSTAT. When possible, we kept the matches for which the location matched better (e.g. full as against 3-digit postal code and street and city as against city city alone), and we removed all the remaining ambiguous matches. Exact matches with coinciding location resulted in 42,376 matched patent applications. Finally, we matched all the non-duplicate exact matches for which the location was different in PATSTAT and AIDA (3,510 applications).

Next we performed approximate matching. To identify possible matches, we used the Perl application published on the new website of the NBER Patent Data Project.[25] Once we had identified candidate approximate matches (some 60,000 pairs), we created various indicators of string similarity.

A first measure is the simple Levenshtein distance (Levensthtein, 1966)[26] between harmonized names in AIDA and PATSTAT for candidate matches. The Levenshtein distance computes the number of single operations (deletion of a character, insertion of a character, substitution of a character and displacement of a character) needed to transform one string into another.[27] Especially when comparing long strings (or long with short strings), a relative measure is more appropriate, so we also consider the ratio between the Levinshtein distance and the maximum or minimum length (in terms of number of characters) of the two strings. Finally, we computed another measure to allow for the possibility that some unnecessary substring had been added in one of the two strings. This measure is given by the difference between the Levenshtein distance and the absolute value of the difference between the length of the two strings $(LEV(A,B) - |length(A) - length(B)|)$.[28]. Finally, we identified all cases in which one of the two strings represented a substring of the other string[29]

After that, we ranked the candidate pairs according too various combinations of measures of string similarity and proceeded to visual identification of the matches, taking a very conservative approach.[30] Especially in the case of EPO applicants, we also ranked candidate matches according to similarity of address (when available). Approximate matching, overall, produces matches for 1,704 patent applications (location coincided for 1,283).

Finally, we added all Thoma et al. (2010) matches for Italian firms, removing the applications that had been matched previously to correct for the fact that Thoma et al. (2010) assigned patent applications of subsidiaries to their parent companies. These matches involved 14,226 EPO applications. The nine steps resulted in the matching of 8,892 EPO applicants and 49,369 applications, divided into four broad categories:[31]

- Exact matches of firms for which the address of the applicant and the address in AIDA somehow coincides (42,376 EPO applications);
- Exact matches of non-duplicate firms with non-coinciding location (3,510 EPO applications);

[24] In some cases the problem of duplicates might be very severe, as in the case of firms whose name is FUTURA SRL (60 occurrences), SIRIO SRL (46 occurrences) and PEGASO SRL (45 occurrences).

[25] As an alternative, we used the RECLINK user-written Stata command (Blasnik, 2007), but it was outperformed by the Perl application both in speed and in effectiveness.

[26] The computation of the Levinsthein distance in Stata was done with the user-written module LEVENSHTEIN (Reif, 2010)

[27] To transform TABLE into CABLE there is need of just one operation (substitution of C for T) which corresponds to a Levinshtein distance of 1. To transform TABLE into CATTLE three operations are needed: substitution of C for T, substitution of T for B, insertion of a T before the L.

[28] In this case, MARIO ROSSI SPA and MARIO ROSSI SPA DI M ROSSI (DI M ROSSI being an unnecessary substring) will have a score of 0 while MARIO ROSSI SPA and MARIA ROSSI SPA will have a score of 1.

[29] MARIO ROSSI SPA is a substring of MARIO ROSSI SPA DI M ROSSI.

[30] The combination of measures of string similarity was changed in every case in which, scrolling the list down to lower level of similarity, no match was found for about 50 pairs. The procedure stopped when all possible combinations of measures had been analysed.

[31] Note that the sum of applicants and applications, when divided, is greater than the aggregate figure because several matches pertain to more than one category

- Approximate matches (1,704 EPO applications);
- BvD codes and publication numbers for EPO applications already matched by Thoma et al. (2010) (14,226 EPO applications).[32]

The first two categories are expected to be the most reliable with small share of false matches; approximate matching is the category for which the share of false positives should be greatest. False matches in the first two categories could occur because of failure to identify duplicates in the list of AIDA firms or because the applicant in PATSTAT corresponds to a firm that is not in the AIDA database. The version of AIDA employed here in fact is a small non-random sample of the larger population of Italian firms, so it is possible that not all the duplicates will be detected.

Each application/applicant pair was tagged by category of match. Pairs reported in multiple categories were assigned to the one with the greatest presumed reliability.[33] Information on the source of matching could be useful in empirical analysis when chosing between the largest sample (with moderate probability of false matches) and the smalle rsample, with the smallest expected share of false matches.

3 Results

3.1 The PATSTAT/AIDA matching

For the period 2000-2007,[34] the matching between firms in AIDA with EPO applicants found 5,485 EPO applicants and 23,501 EPO applications. In addition, we matched 5,008 EPO applicants and 24,120 EPO applications for the period 1977-1999. Given that the version of AIDA used covers only 2000-2009, the matching for 1977-1999 is likely to miss many applicants who exited the market before 2000, but data on patent applications for the earlier period could be useful in creating stock measures at firm level.

To estimate our degree of coverage of the population of Italian firms, we first computed the number of EPO applications with names containing the strings 'SRL', 'SPA', 'SNC', 'SAS', 'SAPA' and 'COOP',[35] which denote the legal status of all Italian companies. The general coverage (i.e., including all matches) is shown in Figure 1.

On average, we were able to match more than 80 percent of the EPO applications filed by Italian firms (82.8 percent for the period 1977-2009, with a peak of 91 percent in 2002).

Figure 1 reports different measures of coverage, combining information of firms in AIDA, with balance-sheet information `non-missing` matched with applicants in PATSTAT with either overall Italian patent applicants (`total`) or 'corporate' ones [36] (`firms`) in PATSTAT.

Table 3 reports some additional information on the coverage for 2000-2007. Name harmonization reduces in the number of applicants in PATSTAT of 5 percent. Matched applicants account for 46 percent of total applicants in PATSTAT and 73 percent of Italian applications to EPO. Finally, there is a small difference (about 1.4 percent) between the total EPO application count by Italian applicants resulting from raw data in PATSTAT (32,203) and the official OECD data (based on PATSTAT, 31,743). This is probably due to the fractional count of applications with applicants resident in two different countries used by OECD aggregate figures.

3.2 Some preliminary descriptive evidence

We can analyze the descriptive evidence arising from our matched dataset, in search of patenting patterns. We focus on the distribution of applicants and applications by time, sector, firm size, location and technology. The results below are restricted to the sub-sample of firm-year pairs for which at least the

[32] When sorting patent applications by priority date for 1990-2003, Thoma et al. (2010) identified 10,240 EPO patent applications; our procedure identified an additional 16,952, a gain of 166 percent.

[33] The rank is: (i) exact match of firms with the same location; (ii) exact match of firms with different addresses; (iii) approximate match; (iv) Thoma's match.

[34] We do not consider applications in 2008 and 2009, because their level is much lower than in earlier years. This is because of the familiar truncation of patent data owing to lag between the application and its publication (Hall et al., 2001). Matched applications for 2008 and 2009 number 2,616.

[35] These strings were searched from the list of harmonized names.

[36] 'Corporate' applicants where identified by looking for any of the following strings in the applicant name in PATSTAT: 'SRL', 'SPA', 'SNC', 'SAS', 'SAPA' and 'COOP'.

book value of the firm could be retrieved. Moreover, most of the statistics were computed on the sub-sample of manufacturing firms, which is justified by the fact that the innovations generally covered by IPRs are, most of the times, in product or process innovations that will be commercially exploited by the manufacturing sector and most of these innovations are created within the manufacturing sector.

We report transition matrices for all sectors, for manufacturing firms and for manufacturing firms in medium-high and high-technology sectors[37] (refer to Tables 5, 6 and 7). Each row describes the distribution of the number of patent applications per firm at *t+1* for firms that were in the size class identified by the row at time *t*. Transition matrices concisely portray the persistence of phenomena. They have been used in two recent articles on patent data (Hingley and Bas, 2009; Helmers et al., 2011) and indicate that persistence in patent applications is increasing in the extent of patenting activity in the past and that many firms file for only one patent during their active life.

First, very few firm-year pairs correspond to positive patent applications (0.7 percent for all EPO applications, 3.67 percent for those in medium-high and high-technology manufacturing sectors). The proportion increases as we move from all firms to manufacturing firms and then to medium-high and high-technology manufacturing. Second, all firms that file more than 20 applications in a year also patent at least 2 EPO applications the following year (about 85 percent of these firms submit 11 or more EPO applications the following year). Finally, firms with no patent applications have a very low probability of filing for 6 or more patents next year (always below 0.02 percent). This last pattern strongly suggests that becoming a large-scale innovator is a cumulative, long-run process.

Comparing our results for aggregate Italian EPO applications with those of Helmers et al. (2011) on EPO and IPO patent applications (see Table 4), we observe that the lower-right part of the transition matrix (firm-year pairs with 2 or more patent applications) is almost identical while the probabilities of transition for firms with at most one or patents differ substantially; the Italian firms with only one application per year are less persistent. This suggests that once the hurdle of filing is passed, innovative Italian firms behave similarly to their European counterparts (Lotti and Schivardi (2005)).

Table 8 shows the propensity to patent in different macro-sectors expressed, as share of firm-year pairs with at least one application. R&D services and manufacturing (especially medium-high and high-technology manufacturing) tend to patent more. Of course, patents are an output measure for firms in the R&D sector. Their innovations are generally transferred and licensed to firms in the industrial (especially manufacturing) sector. Patents allow firms in the R&D sector to appropriate of part of the commercial value of the innovations. The other service sectors have very little propensity to patent, as do the sectors of electricity, gas and water, construction, and agriculture and mining sectors.

Table 9 shows the sectoral distribution of patent applications, table 10 the relative contribution of each sector. Manufacturing sectors alone account for about 81 percent of EPO applications, while the R&D sector, despite its very high propensity to patent, contributes little to the total patents, reflecting the small number of R&D firms. As expected, within manufacturing medium-high and high-technology sectors are characterized by greater propensity to patent and a larger contribution to total patenting activity.

The propensity to patent and the distribution of patent applications in manufacturing alone are given in Tables 11 and 12. The propensity to patent is regularly above 2 percent for seven sectors, all medium-high or high-technology (Nace Rev. 1.1 codes 24, 25, 29, 31, 32, 33 and 34). Most of the other manufacturing sectors show very little propensity to patent, generally below 1 percent. The distribution of sectoral contributions to total manufacturing patent applications is even more skewed, with five sectors (Nace Rev. 1.1 codes 29, 24, 32, 28 and 31) accounting for about 70 percent.

Tables 13 and 14 and Figures 3 and 4 report the distribution by macro-region and province of applications and applicants.[38] Patent propensity of manufacturing firms is much greater in the northern regions (2 to 2.6 percent) than in central Italy (1 to 1.4 percent) or the South (0.3 to 0.5 percent). The same pattern characterizes regional contribution to aggregate manufacturing applications (86 percent of EPO applications filed by firms in the North and just 1-2 percent by the southern firms). This geographical

[37] Medium-high and high-technology firms are, following the OECD definition, those firms in following Nace Rev. 1.1 sectors: 23 (coke, refined petroleum products and nuclear fuel), 24 (chemicals and chemical products), 25 (rubber and plastic products), 26 (other non-metallic mineral products), 27 (basic metals), 28 (fabricated metal products, except machinery and equipment), 29 (machinery and equipment n.e.c.), 30 (office machinery and computers), 31 (electrical machinery and apparatus n.e.c.), 32 (radio, television and communication equipment and apparatus), 33 (medical, precision and optical instruments, watches and clocks), 34 (motor vehicles, trailers and semi-trailers) and 35 (other transport equipment).

[38] Also in this case we restrict examination to the firm/year pairs for which balance-sheet information was available.

concentration presumably reflects differences in the sectoral mix, in local systems of innovation and in endowments of physical and human capital.

The provincial distribution of patenting activity by province (here the results refer to all sectors) is marked by strong concentration of applications in just a few areas: provinces of Lombardy and Veneto between Milan and Venice, the provinces of Emilia-Romagna between Bologna and Piacenza, Turin and Rome. No southern province has significant patenting activity. Patenting itself follows a similar pattern, the only notable difference being low propensity in Rome and high patent propensity in the Marche region and in the provinces of Chieti (Abruzzo) and Isernia (Molise).

Tables 15 and 16 report patent propensity and the distribution of manufacturing patent applications by firm size.[39] The contribution of large firms to total applications is very large (between 50 and 60 percent); more generally, as expected, both patent propensity and relative contribution to total applications is decrease as firm size diminishes.

Finally, let us discuss the distribution of EPO applications by technology class (Table 17). We classified them according to various technology classifications derived from IPC classes. First, we used the ISI-OST-INPI classification (8th edition 2006, Schmoch (2008)), which groups hundreds of thousands of IPC classes into 30 or 7 macro-areas. The upper part of Table 17 reports the classification in 7 macro-areas.[40] The distribution according to technology is extremely persistent, with no significant shift in the period considered. The two most common fields are in mechanical engineering, machines, transport and industrial processes, account for more than a quarter of all EPO applications in manufacturing. Pharmaceutical and biotechnology account for just 8 percent of manufacturing patents, but they are highly concentrated in the pharmaceutical sector. Second, we identified environmental patents according to two different selections of environmental IPC classes: the 'IPC Green Inventory'[41] created by the World Intellectual Property Organization and the United Nations Framework Convention on Climate Change, covering Environmentally Sound Technologies and the 'Series of patent search strategies for the identification of selected environment-related technologies'[42] developed by the OECD. The OECD's approach is more restrictive than that of the WIPO (which includes most of the environmental patents already identified by the OECD). The share of environmental patents is quite low (about 8 percent in the aggregate and 3 percent as regards those identified by the OECD) and does not show any clear trend, either upward or downward. Finally, ICT patents (as defined by the OECD[43] according to their IPC class) represent a significant though decreasing share of total EPO applications by manufacturing, shrinking from 21 percent in 2000 to 12 percent in 2007.

4 Conclusion

This paper describes the method used to create an integrated base of data on firms' financial accounts and patent applications in Italy. The administrative nature of the financial accounts and the relevance and flexibility of patents as a measure of innovation output will enable researchers to inquire into a series of issues involving patterns of innovation at the firm level. It goes without saying that in this field one must always bear in mind the potential selection biases and measurement errors due to false positive and negative matches and errors in the financial accounts.

This integrated database can be easily extended with the rich information contained in patent data such as citation links, patent families, PCT applications and information on inventors.

[39] The European Commission (Recommendation 2003/361/EC), defines macro classes of firms as follows: (i) micro firms: fewer than 10 employees and turnover or book value of less than 2 million euros; (ii) small firms: 11-50 employees and a turnover or book value of 2 to 10 million euros; (iii) medium-sized firms: 51-250 employees and a turnover between or book value of 10-50 million euros; (iv) large firms: residually those not included in the foregoing classes.

[40] The percentages do not sum up to 100 because several applications contain multiple IPC classes pertaining to different ISI-OST-INPI categories.

[41] http://www.wipo.int/classifications/ipc/en/est/

[42] http://www.oecd.org/dataoecd/4/14/47917636.pdf

[43] http://www.oecd.org/dataoecd/34/34/40807441.pdf

A Appendix

Table 1: Number of firms with non-missing balance-sheet information

	2000	2001	2002	2003	2004	2005	2006	2007	Total
Agric. Mining	1721	2300	3063	3225	3888	4103	4238	4423	26961
MH-tech manuf	10626	12075	14688	15179	17544	18426	19091	19915	127544
Low-tech manuf	23905	27268	33148	34179	39897	42223	44009	46271	290900
EGW, construction	7785	9780	14807	16054	23438	26366	29127	33276	160633
Wholesale, retail, hotel	27489	32670	43103	45110	56263	60862	64865	69969	400331
Transport and telecom	4673	5551	7437	7932	10224	11184	11811	12750	71562
Finance, real estate	2503	3571	4221	5504	11805	13431	14967	17584	73586
Computer	1449	1902	3100	3325	4411	4718	4905	5173	28983
R&D services	134	174	263	281	363	386	417	461	2479
Business activities	3345	4875	7954	8757	12036	13200	14460	15818	80445
Other services	2512	3533	5691	6171	8209	8778	9237	9852	53983
Total	86142	103699	137475	145717	188078	203677	217127	235492	1317407

Source: AIDA database.

Table 2: Number of firms with non-missing balance-sheet information (share of total)

	2000	2001	2002	2003	2004	2005	2006	2007	Total
Agric. Mining	2.0	2.2	2.2	2.2	2.1	2.0	2.0	1.9	2.0
MH-tech manuf	12.3	11.6	10.7	10.4	9.3	9.0	8.8	8.5	9.7
Low-tech manuf	27.8	26.3	24.1	23.5	21.2	20.7	20.3	19.6	22.1
EGW, construction	9.0	9.4	10.8	11.0	12.5	12.9	13.4	14.1	12.2
Wholesale, retail, hotel	31.9	31.5	31.4	31.0	29.9	29.9	29.9	29.7	30.4
Transport and telecom	5.4	5.4	5.4	5.4	5.4	5.5	5.4	5.4	5.4
Finance, real estate	2.9	3.4	3.1	3.8	6.3	6.6	6.9	7.5	5.6
Computer	1.7	1.8	2.3	2.3	2.3	2.3	2.3	2.2	2.2
R&D services	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
Business activities	3.9	4.7	5.8	6.0	6.4	6.5	6.7	6.7	6.1
Other services	2.9	3.4	4.1	4.2	4.4	4.3	4.3	4.2	4.1
Total	100	100	100	100	100	100	100	100	100

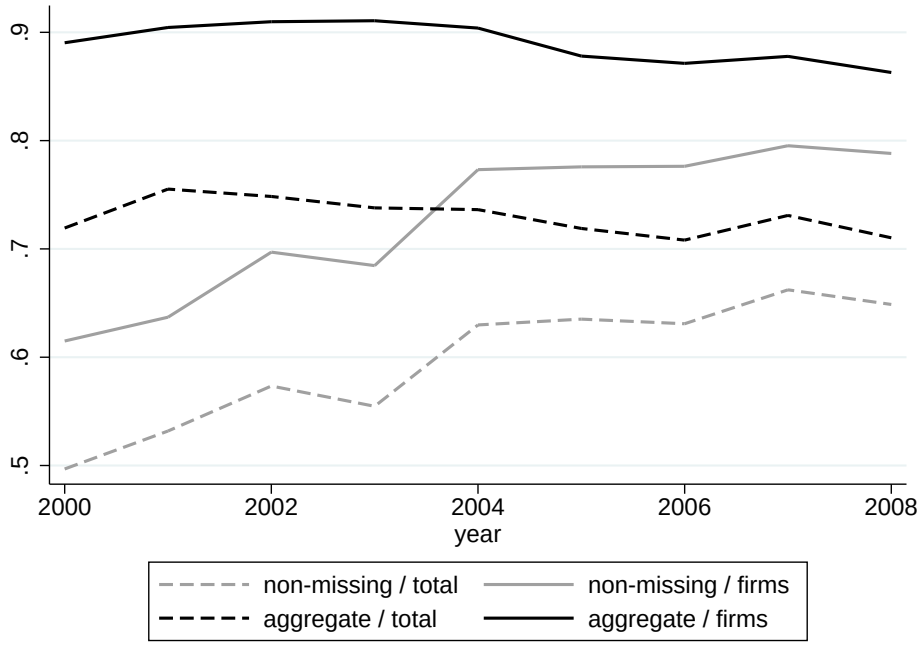
Source: AIDA database.

Table 3: Some figures on the coverage of the matching (2000-2007)

Number of applicants in Patstat before name harmonization	12732
Number of applicants in Patstat after name harmonization	12054
Number of applicants matched	5485
Number of EPO applications by Italian applicants (OECD)	31743
Number of EPO applications by Italian applicants (own elaboration on Patstat)	32203
Number of matched applications	20501

Source: Matched AIDA-PATSTAT database and PATSTAT database.

Figure 1: Coverage (%) of the AIDA/PATSTAT matching



Source: Matched AIDA-PATSTAT database and PATSTAT database. *Non-missing*: firm/year pair with non-missing balance-sheet information; *Aggregate*: includes also firm/year pairs with missing balance-sheet; *Firms*: patent applications in Patstat by 'corporations' (applicants containing the strings 'SRL', 'SPA', 'SNC', 'SAS', 'SAPA' or 'COOP'); *Total*: patent applications in Patstat by all applicants.

Table 4: Transition matrix for applications at EPO and UK Patent Office for firms in FAME taken from Helmers et al. (2011) (all sectors - in %)

	No patent	1 patent	2-5 patents	6-10 patents	11-20 patents	20+ patents	Total
No patent	80.66	16.35	2.87	0.09	0.03	0	100
1 patent	71.24	19.92	8.23	0.44	0.13	0.03	100
2-5 patents	40.75	26.82	26.89	4.91	0.67	0.07	100
6-10 patents	7.6	15.2	37.22	26.61	11.66	1.7	100
11-20 patents	3.63	3.3	20.46	29.37	30.69	12.54	100
20+ patents	1.04	1.55	1.04	5.7	22.28	68.39	100
Total	75.99	17.48	5.39	0.68	0.29	0.17	100

Source: Helmers et al. (2011).

Table 5: Transition matrix (EPO applications - all sectors - in %)

	No patent	1 patent	2-5 patents	6-10 patents	11-20 patents	20+ patents	Total
No patent	99.55	0.35	0.09	0	0	0	100
1 patent	74.57	16.35	8.55	0.47	0.06	0	100
2-5 patents	45.47	21.55	26.84	5.09	0.86	0.2	100
6-10 patents	7.39	9.13	39.57	27.39	15.22	1.3	100
11-20 patents	0.86	5.17	15.52	25.86	41.38	11.21	100
20+ patents	0	0	5.56	8.33	19.44	66.67	100
Total	99.3	0.47	0.19	0.02	0.01	0.01	100

Source: Matched AIDA-PATSTAT database.

Table 6: Transition matrix (EPO applications - manufacturing - in %)

	No patent	1 patent	2-5 patents	6-10 patents	11-20 patents	20+ patents	Total
No patent	98.78	0.95	0.26	0.01	0	0	100
1 patent	74.06	16.22	9.17	0.55	0	0	100
2-5 patents	44.96	22.12	26.4	5.37	0.87	0.29	100
6-10 patents	8.43	4.82	42.17	26.51	16.27	1.81	100
11-20 patents	0	6.25	17.5	21.25	43.75	11.25	100
20+ patents	0	0	4.35	15.22	19.57	60.87	100
Total	98.11	1.25	0.53	0.07	0.03	0.02	100

Source: Matched AIDA-PATSTAT database.

Table 7: Transition matrix (EPO applications - medium-high and high-tech manufacturing - in %)

	No patent	1 patent	2-5 patents	6-10 patents	11-20 patents	20+ patents	Total
No patent	97.76	1.71	0.5	0.02	0	0	100
1 patent	71.74	17.56	10.08	0.62	0	0	100
2-5 patents	41.92	22.12	27.77	6.64	1.33	0.22	100
6-10 patents	6.8	4.76	41.5	28.57	17.01	1.36	100
11-20 patents	0	7.25	13.04	24.64	44.93	10.14	100
20+ patents	0	0	3.03	15.15	21.21	60.61	100
Total	96.33	2.29	1.08	0.18	0.09	0.04	100

Source: Matched AIDA-PATSTAT database.

Table 8: % of firms with at least one EPO application

	2000	2001	2002	2003	2004	2005	2006	2007	Total
Agric. Mining	0.06	0.00	0.03	0.06	0.10	0.05	0.07	0.07	0.06
MH-tech manuf	3.60	3.69	3.53	3.49	3.32	3.44	3.35	3.32	3.44
Low-tech manuf	0.94	0.94	0.94	0.91	0.87	0.97	0.91	0.88	0.92
EGW, construction	0.17	0.16	0.14	0.13	0.14	0.14	0.11	0.15	0.14
Wholesale, retail, hotel	0.20	0.15	0.18	0.16	0.14	0.12	0.13	0.14	0.15
Transport and telecom	0.02	0.13	0.07	0.04	0.08	0.06	0.11	0.07	0.07
Finance, real estate	0.44	0.39	0.33	0.24	0.13	0.13	0.09	0.10	0.16
Computer	0.21	0.26	0.13	0.24	0.25	0.23	0.16	0.25	0.22
R&D services	8.21	6.90	5.70	6.05	5.51	5.96	7.19	4.99	6.09
Business activities	0.63	0.68	0.50	0.41	0.34	0.38	0.31	0.35	0.40
Other services	0.04	0.14	0.05	0.08	0.13	0.09	0.11	0.11	0.10
Total	0.84	0.81	0.73	0.70	0.61	0.62	0.59	0.57	0.66

Source: Matched AIDA-PATSTAT database.

Table 9: Number of EPO applications

	2000	2001	2002	2003	2004	2005	2006	2007	Total
Agric. Mining	1	0	2	2	5	2	3	3	18
MH-tech manuf	1138	1164	1434	1354	1525	1664	1649	1656	11584
Low-tech manuf	396	463	499	536	577	661	666	654	4452
EGW, construction	19	21	34	38	50	52	51	88	353
Wholesale, retail, hotel	76	108	114	117	106	105	106	135	867
Transport and telecom	1	45	71	85	103	95	74	52	526
Finance, real estate	35	45	57	51	58	35	41	40	362
Computer	3	6	6	10	15	17	9	14	80
R&D services	62	94	71	110	127	73	90	50	677
Business activities	39	54	58	76	84	113	94	106	624
Other services	2	11	6	10	15	13	18	21	96
Total	1772	2011	2352	2389	2665	2830	2801	2819	19639

Source: Matched AIDA-PATSTAT database.

Table 10: Number of EPO applications (% of total)

	2000	2001	2002	2003	2004	2005	2006	2007	Total
Agric. Mining	0.06	0.00	0.09	0.08	0.19	0.07	0.11	0.11	0.09
MH-tech manuf	64.22	57.88	60.97	56.68	57.22	58.80	58.87	58.74	58.98
Low-tech manuf	22.35	23.02	21.22	22.44	21.65	23.36	23.78	23.20	22.67
EGW, construction	1.07	1.04	1.45	1.59	1.88	1.84	1.82	3.12	1.80
Wholesale, retail, hotel	4.29	5.37	4.85	4.90	3.98	3.71	3.78	4.79	4.41
Transport and telecom	0.06	2.24	3.02	3.56	3.86	3.36	2.64	1.84	2.68
Finance, real estate	1.98	2.24	2.42	2.13	2.18	1.24	1.46	1.42	1.84
Computer	0.17	0.30	0.26	0.42	0.56	0.60	0.32	0.50	0.41
R&D services	3.50	4.67	3.02	4.60	4.77	2.58	3.21	1.77	3.45
Business activities	2.20	2.69	2.47	3.18	3.15	3.99	3.36	3.76	3.18
Other services	0.11	0.55	0.26	0.42	0.56	0.46	0.64	0.74	0.49
Total	100	100	100	100	100	100	100	100	100

Source: Matched AIDA-PATSTAT database.

Table 11: % of firms with at least one EPO application (manufacturing - NACE Rev. 1.1)

	2000	2001	2002	2003	2004	2005	2006	2007	Total
15	0.20	0.28	0.49	0.39	0.27	0.29	0.27	0.34	0.32
16	0.00	0.00	0.00	0.00	0.00	6.67	0.00	0.00	0.86
17	0.29	0.65	0.50	0.68	0.55	0.92	0.58	0.59	0.61
18	0.38	0.16	0.13	0.06	0.21	0.29	0.28	0.35	0.24
19	0.61	0.52	0.44	0.42	0.61	0.66	0.50	0.46	0.53
20	0.00	0.31	0.32	0.23	0.06	0.24	0.17	0.42	0.23
21	0.60	1.19	1.12	0.92	1.14	0.86	1.45	1.20	1.08
22	0.14	0.24	0.05	0.18	0.26	0.29	0.10	0.10	0.17
23	0.69	1.25	1.13	0.56	1.03	0.00	0.00	0.50	0.62
24	3.83	3.85	3.53	4.15	3.78	3.78	3.33	2.99	3.63
25	2.60	2.64	2.39	2.39	2.66	2.99	2.63	2.29	2.58
26	0.77	0.74	0.51	0.63	0.49	0.67	0.59	0.59	0.61
27	1.54	0.71	1.33	1.03	1.50	1.06	0.82	1.25	1.15
28	1.46	1.34	1.33	1.48	1.30	1.39	1.46	1.30	1.38
29	4.08	3.97	4.04	3.82	3.59	3.85	3.58	3.71	3.80
30	2.03	1.72	1.23	1.73	1.17	2.62	2.35	2.53	1.98
31	2.96	2.72	2.75	2.86	2.81	2.77	2.81	2.72	2.79
32	2.19	3.79	2.95	2.97	2.75	2.40	2.31	3.12	2.79
33	3.28	4.21	3.64	3.21	3.53	3.71	4.39	3.83	3.75
34	4.00	4.70	3.52	3.81	3.92	3.44	4.05	3.88	3.89
35	1.62	1.61	1.99	1.49	1.63	1.40	1.81	1.71	1.66
36	1.26	1.14	1.49	0.86	0.73	0.93	0.93	0.96	1.01
37	0.00	0.00	0.00	0.00	0.19	0.18	0.00	0.00	0.06
Total	1.44	1.42	1.32	1.29	1.15	1.20	1.13	1.07	1.22

Source: Matched AIDA-PATSTAT database.

Table 12: Number of EPO applications (% of total - manufacturing - NACE Rev. 1.1)

	2000	2001	2002	2003	2004	2005	2006	2007	Total
15	0.64	1.09	1.63	1.76	1.35	1.18	1.39	1.00	1.27
16	-	-	-	-	-	0.13	-	-	0.02
17	0.39	1.15	0.71	1.09	0.88	1.47	1.01	1.46	1.06
18	0.26	0.18	0.15	0.05	0.46	0.42	0.59	0.79	0.39
19	0.52	0.42	0.36	0.47	0.60	0.67	0.68	0.58	0.55
20	-	0.18	0.20	0.21	0.05	0.29	0.21	0.38	0.20
21	0.77	1.15	0.86	0.73	0.88	1.18	1.39	0.92	1.00
22	0.13	0.30	0.05	0.21	0.33	0.46	0.13	0.17	0.23
23	0.06	0.12	0.10	0.05	0.09	-	-	0.04	0.05
24	12.69	13.71	16.07	16.55	14.45	13.04	12.00	10.09	13.45
25	6.83	7.22	5.39	6.43	6.64	5.89	6.76	5.59	6.30
26	1.67	1.52	1.02	1.19	1.02	1.56	1.52	1.63	1.39
27	1.35	0.67	0.97	0.93	1.16	1.26	0.80	0.96	1.01
28	9.08	10.98	10.37	12.34	11.48	11.11	11.37	11.22	11.06
29	28.85	27.00	28.52	26.30	28.16	27.98	28.53	28.69	28.04
30	0.52	0.24	0.20	0.36	0.23	0.55	0.55	0.83	0.45
31	6.57	7.89	7.32	7.57	6.74	7.99	8.50	10.34	7.97
32	14.17	11.35	11.03	10.06	10.22	9.93	7.86	6.38	9.84
33	4.12	4.67	4.98	3.73	4.32	4.42	5.79	4.63	4.62
34	5.80	4.98	3.86	4.93	4.88	3.87	4.23	5.21	4.67
35	0.58	0.79	0.92	0.73	1.86	2.23	2.24	2.88	1.64
36	3.80	3.09	3.56	2.33	1.81	2.15	2.28	2.54	2.62
37	-	-	-	-	0.05	0.04	-	-	0.01
Total	100	100	100	100	100	100	100	100	100

Source: Matched AIDA-PATSTAT database.

Table 13: % of firms with at least one EPO application (manufacturing)

	2000	2001	2002	2003	2004	2005	2006	2007	Total
North-West	2.18	2.08	2.01	1.88	2.02	2.28	2.16	2.10	2.10
North-East	2.03	2.22	2.01	2.09	2.25	2.59	2.44	2.48	2.28
Central Italy	1.06	1.20	1.15	1.38	1.17	1.43	1.16	1.25	1.23
South and islands	0.33	0.50	0.36	0.40	0.36	0.45	0.29	0.43	0.39
Total	1.76	1.82	1.70	1.74	1.76	2.04	1.86	1.88	1.83

Source: Matched AIDA-PATSTAT database.

Table 14: Number of EPO applications (% of total - manufacturing)

	2000	2001	2002	2003	2004	2005	2006	2007	Total
North-West	59.36	53.53	56.94	56.42	50.16	47.35	52.54	49.35	52.92
North-East	28.59	32.37	29.24	29.82	37.15	38.61	34.82	34.87	33.43
Central Italy	10.92	12.88	12.71	12.52	11.06	12.82	11.23	13.57	12.23
South and islands	1.13	1.22	1.11	1.25	1.63	1.23	1.40	2.21	1.42
Total	100	100	100	100	100	100	100	100	100

Source: Matched AIDA-PATSTAT database.

Table 15: % of firms with at least one EPO application (manufacturing)

	2000	2001	2002	2003	2004	2005	2006	2007	Total
Micro firms	0.51	0.39	0.45	0.33	0.40	0.46	0.41	0.42	0.42
Small firms	0.90	0.79	0.87	0.90	0.92	1.15	1.08	1.09	0.98
Medium firms	3.76	3.90	4.01	3.92	4.34	4.91	4.48	4.73	4.28
Large firms	16.51	16.93	16.47	16.49	16.50	17.28	17.01	17.10	16.81
Total	1.76	1.82	1.70	1.74	1.76	2.04	1.86	1.88	1.83

Source: Matched AIDA-PATSTAT database.

Table 16: Number of EPO applications (% of total - manufacturing)

	2000	2001	2002	2003	2004	2005	2006	2007	Total
Micro firms	3.77	2.30	3.75	2.36	2.93	2.79	2.95	3.37	3.04
Small firms	14.16	12.73	15.23	16.12	19.71	20.21	18.35	20.05	17.34
Medium firms	21.71	25.90	25.37	24.38	28.50	29.71	26.78	27.14	26.33
Large firms	60.36	59.06	55.65	57.14	48.86	47.29	51.91	49.45	53.29
Total	100	100	100	100	100	100	100	100	100

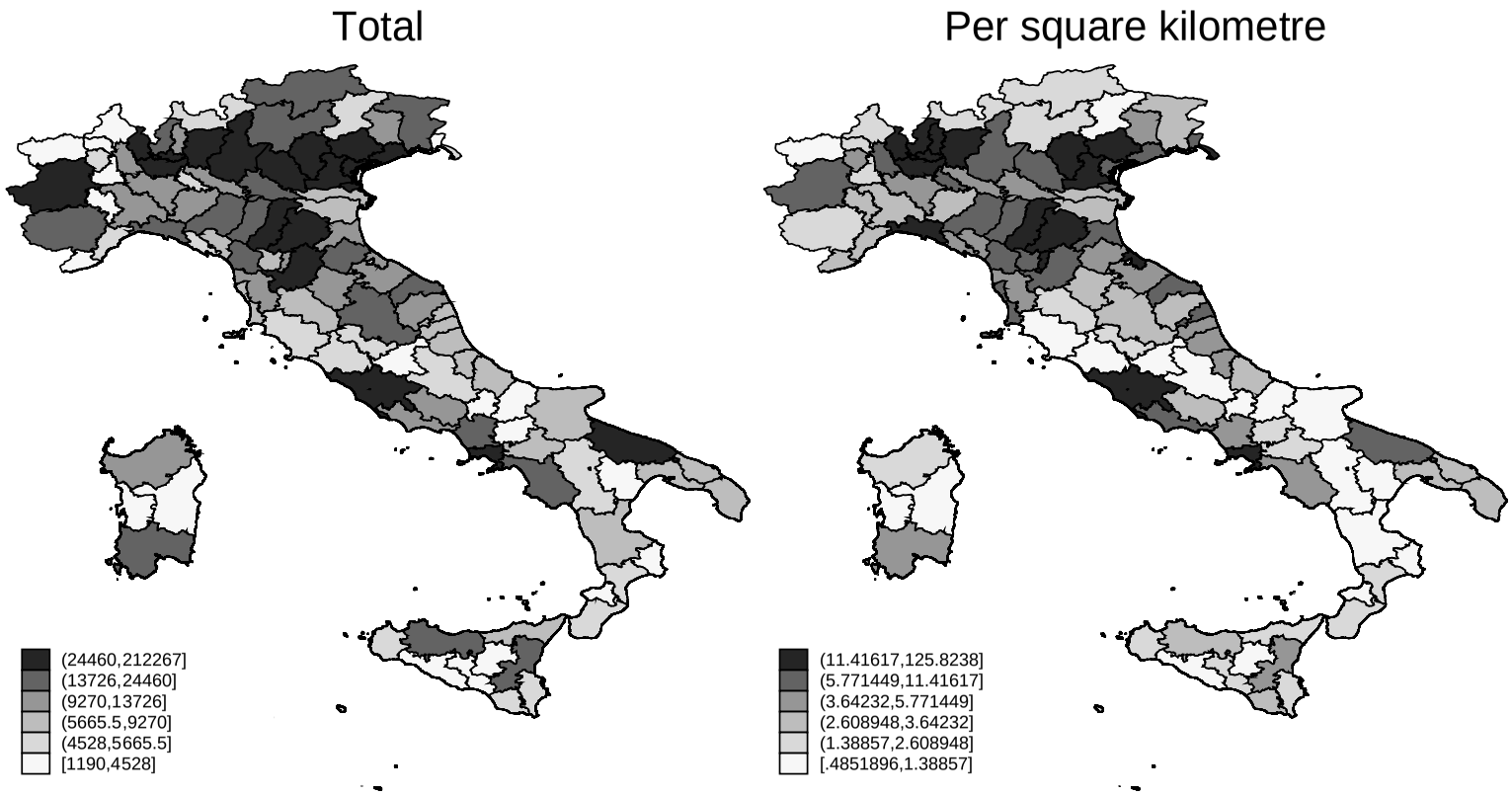
Source: Matched AIDA-PATSTAT database.

Table 17: EPO applications by technology domain (OST7 classification and other classifications - % of total patents - manufacturing)

	2000	2001	2002	2003	2004	2005	2006	2007	Total
Electrical engineering; Electronics	21	16	19	18	11	11	15	15	15
Instruments	13	12	12	11	10	10	12	11	11
Chemicals; Materials	12	12	15	13	13	12	11	9	12
Pharmaceuticals; Biotechnology	7	8	8	7	6	7	5	5	7
Industrial processes	26	31	27	29	32	29	26	25	28
Mechanical eng.; Machines; Transport	30	28	26	29	30	29	30	29	29
Consumer goods; Civil engineering	15	15	19	16	21	21	19	22	19
Environmental patents (OECD)	2	3	2	3	4	3	2	3	3
Environmental patents (WIPO)	7	7	5	7	7	7	7	8	7
Environmental patents (OECD+WIPO)	7	8	6	8	7	8	7	9	8
ICT patents (OECD)	21	17	17	18	9	8	13	12	14

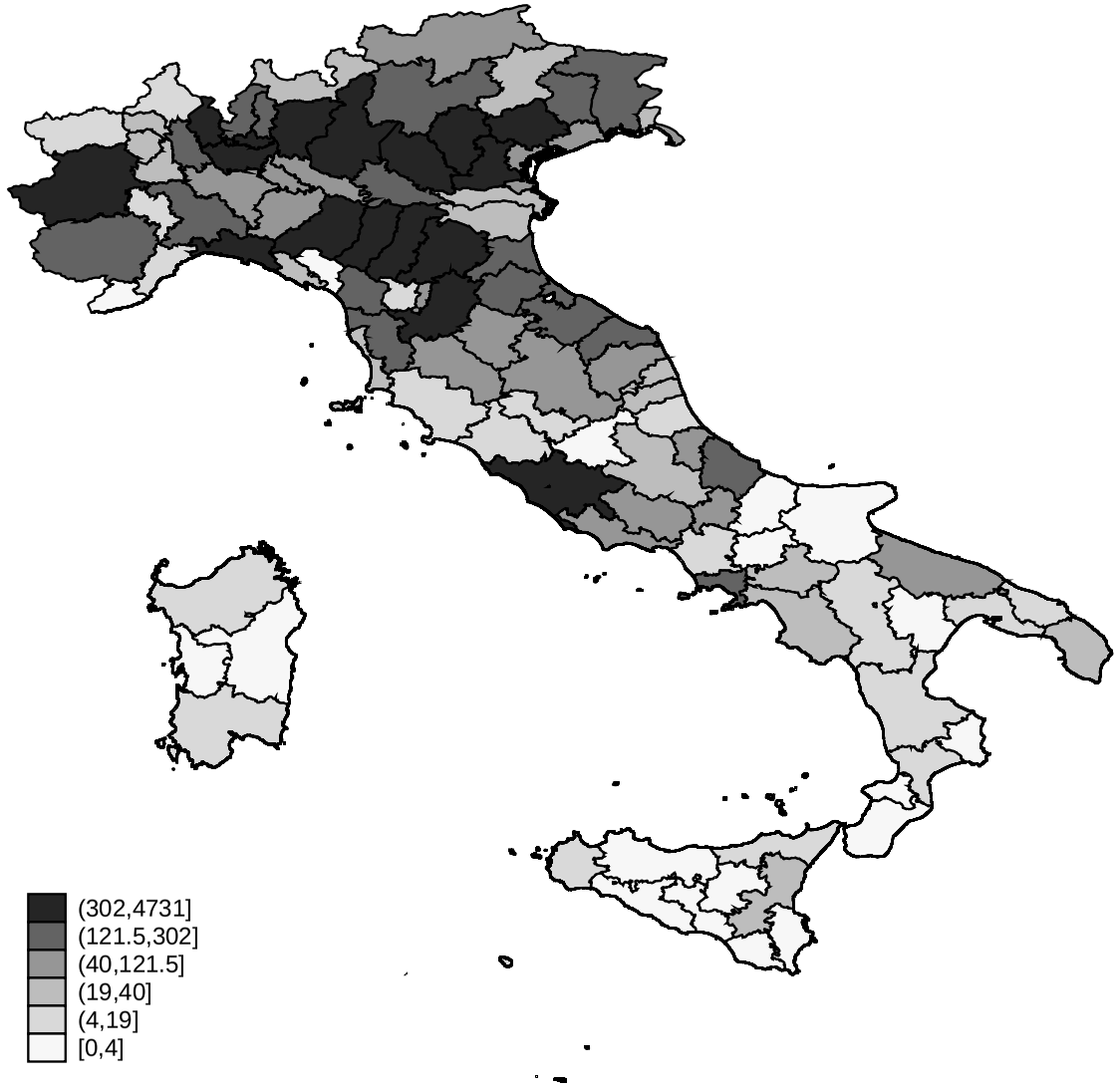
Source: Matched AIDA-PATSTAT database.

Figure 2: Number of firm / year pairs by province (period 2000-2007 - only firms with balance-sheet information)



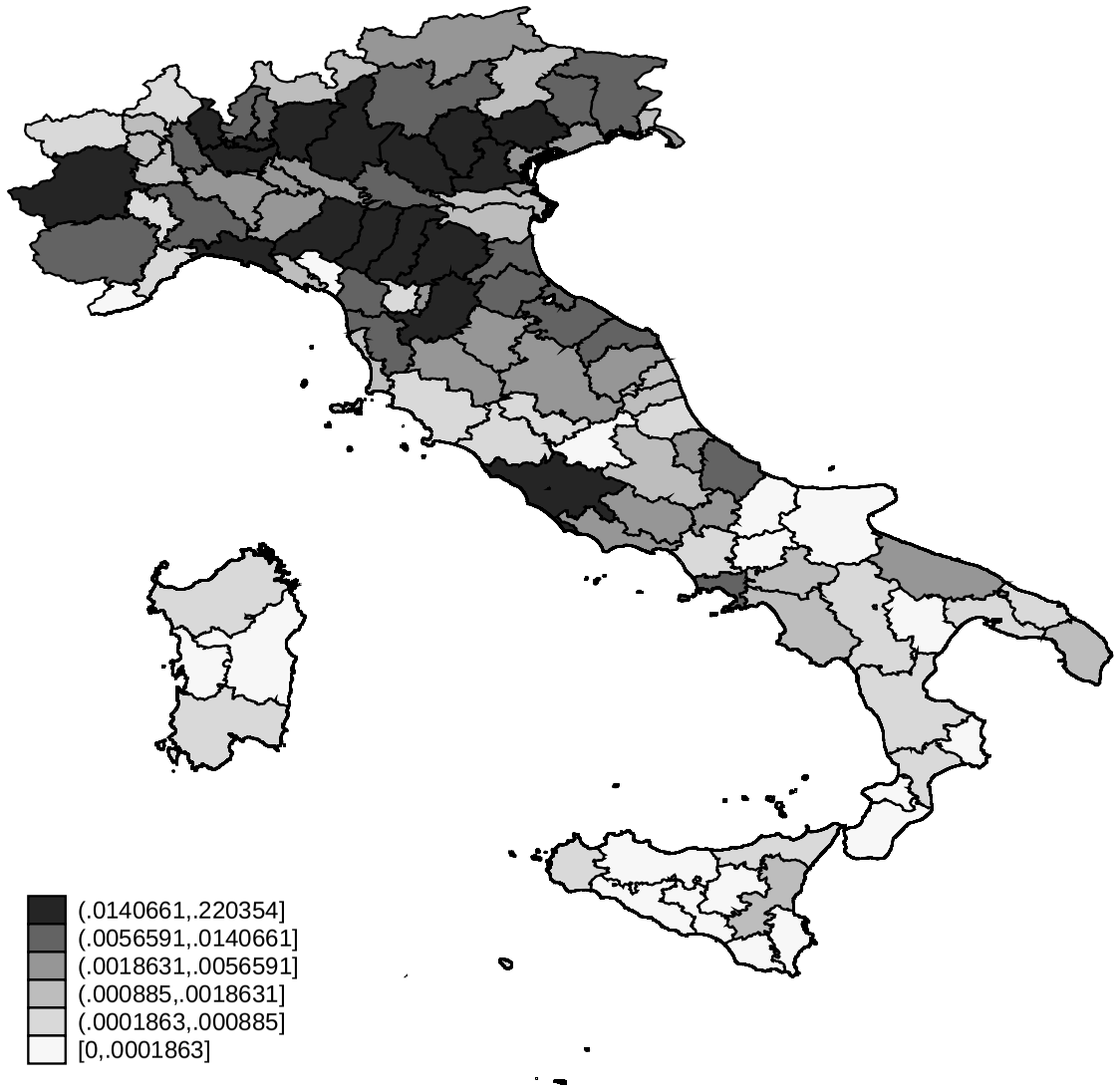
Source: AIDA database.

Figure 3: Number of patent applications by province (period 2000-2007 - only firms with balance-sheet information)



Source: Matched AIDA-PATSTAT database.

Figure 4: Share of firms with at least one patent application by province (period 2000-2007 - only firms with balance-sheet information)



Source: Matched AIDA-PATSTAT database.

References

Blasnik M. (2007), "RECLINK: Stata module to probabilistically match records." Statistical Software Components, Boston College Department of Economics

Bound J., Cummins C., Griliches Z., Hall B. H., and Jaffe A. B. (1984), "Who does R&D and who patents?" In "R&D, Patents, and Productivity" NBER Chapters (National Bureau of Economic Research) pp. 21–54

Cockburn I. M. A., Agrawal A., Bessen J., Graham J. H. S., Hall B. H., and MacGarvie M. (2009), "The NBER patent citations datafile updated." Technical Report

Comanor W. S., Scherer F. M. (1969), "Patent statistics as a measure of technical change." *Journal of Political Economy* 77(3), 392–98

Crepon B., Duguet E., and Mairesse J. (1998), "Research, innovation, and productivity: An econometric analysis at the firm level." NBER Working Papers 6696, National Bureau of Economic Research

De Rassenfosse G., Wastyn A. (2012), "Selection bias in innovation studies: A simple test." ZEW Discussion Papers 12-012, ZEW - Zentrum fr Europische Wirtschaftsforschung / Center for European Economic Research

Griffith R., Huergo E., Mairesse J., and Peters B. (2006), "Innovation and productivity across four European countries." *Oxford Review of Economic Policy* 22(4), 483–498

Griliches Z. (1990), "Patent statistics as economic indicators: A survey." *Journal of Economic Literature* 28(4), 1661–1707

Griliches Z., Pakes A., and Hall B. H. (1988), "The value of patents as indicators of inventive activity." NBER Working Papers, National Bureau of Economic Research

Hall B. H., Cumminq C., Laderman E. S., and Mundy J. (1988), "The R&D master file documentation." NBER Technical Working Papers 0072, National Bureau of Economic Research

Hall B. H., Griliches Z., and Hausman J. A. (1986), "Patents and R&D: Is there a lag?" *International Economic Review* 27(2), 265–83

Hall B. H., Jaffe A. B., and Trajtenberg M. (2001), "The NBER Patent Citation Data File: Lessons, insights and methodological tools." NBER Working Papers, National Bureau of Economic Research

Hausman J., Hall B. H., and Griliches Z. (1984), "Econometric models for count data with an application to the patents-R&D relationship." *Econometrica* 52(4), 909–38

Helmers C., Rogers M., and Schautschick P. (2011), "Intellectual property at the firm-level in the UK: The Oxford firm-level intellectual property database." Economics Series Working Papers 546, University of Oxford, Department of Economics

Hingley P., Bas S. (2009), "Numbers and sizes of applicants at the European Patent Office." *World Patent Information* 31(4), 285–298

Levensthtein V. I. (1966), "Binary codes capable of correcting deletions, insertions and reversals." *Soviet Physics Doklady* 10(8), 707–710

Lissoni F., Maurino A., Pezzoni M., and Tarasconi G. (2010), "APE-INV's "name game" algorithm challenge: A guideline for benchmark data analysis and reporting." Technical Report, Academic Patenting in Europe - APE-INV

Lotti F., Schivardi F. (2005), "Cross country differences in patent propensity: A firm-level investigation." *Giornale degli Economisti* 64(4), 469–502

Mairesse J., Mohnen P. (2010), "Using innovations surveys for econometric analysis." NBER Working Papers 15857, National Bureau of Economic Research

OECD (2009), *OECD Patent Statistics Manual* (OECD)

Pakes A., Griliches Z. (1980), "Patents and R&D at the firm level: A first look." NBER Working Papers 0561, National Bureau of Economic Research

Pavitt K. (1984), "Sectoral patterns of technical change: Towards a taxonomy and a theory." *Research Policy* 13(6), 343–373

Pavitt K. (1985), "Patent statistics as indicators of innovative activities: Possibilities and problems." *Scientometrics* 7(1-2), 77–99

Reif J. (2010), "STRGROUP: Stata module to match strings based on their Levenshtein edit distance." Statistical Software Components, Boston College Department of Economics

Schmoch U. (2008), "Concept of a technology classification for country comparisons." Final Report to the World Intellectual Property Organisation, WIPO

Thoma G., Torrisi S. (2007), "Creating powerful indicators for innovation studies with approximate matching algorithms. a test based on PATSTAT and Amadeus databases." KITeS Working Papers 211, KITeS, Centre for Knowledge, Internationalization and Technology Studies, Università Bocconi, Milan, Italy

Thoma G., Torrisi S., Gambardella A., Guellec D., Hall B. H., and Harhoff D. (2010), "Harmonizing and combining large datasets - An application to firm-level patent and accounting data." NBER Working Papers 15851, National Bureau of Economic Research