Integrating registers: Italian business register and patenting enterprises

Daniela Ichim, Giulio Perani, Giovanni Seri
ISTAT, The Italian National Statistical Institute
Via C. Balbo, 16, 00184, Rome, Italy
{ichim,perani,seri}@istat.it

Abstract: The paper describes the record linkage scheme followed at the Italian national statistical institute to match micro-data on patent application from the international database PATSTAT with the data available from the Italian Official Business Register (ASIA).

The target data in PATSTAT are the applicants based in Italy registering patent/s in the period 1985-2010. Patents applicants can be 'individuals' or 'establishments'. In this last category we aim at identifying business enterprises who were active (as recorded in ASIA) in the period 1989-2008. The wishing output of the linkage process is, for each patenting enterprise, a pair composed by the 'applicant identification code in PATSTAT' and the 'enterprise identification number in ASIA'. This last allows for accessing the repositories of the official statistical data and, therefore, linking economic data to patenting enterprises. Statistical analysis such as: identifying the premises of patenting propensity; evaluate the impact of patenting on the enterprise profitability; etc. can be then performed.

On the methodological side, linkage of patent data has to rely on the 'applicants names'. Consequently, a great effort has been put in the pre-processing phase of the process to standardise the applicant/enterprise names and extract the 'legal form' from the name string. During the linkage process, two practical problems were faced: the reduced number of comparison variables and the huge dimension, in terms of number records, of the Italian Business Register. These issues were addressed within a rule-based deterministic record linkage approach. In this paper, together with the results obtained, we will illustrate the main features of the sequential searching and linkage methodology we adopted.

Keywords: patents, business register, deterministic record linkage

1. Introduction

The paper describes the record linkage scheme followed at the Italian national statistical institute (Istat) to match micro-data on patent application from the PATSTAT database with the data available from the Italian Official Business Register (ASIA) as a preliminary stage of a project aiming, mainly, at monitoring and profiling Italian patenting enterprises.

The target data in PATSTAT are the applicants based in Italy registering patent/s in the period 1985-2010. Patents applicants can be 'individuals' or 'establishments'. In this last category we aim at identifying business enterprises who were active (as recorded in ASIA) in the period 1989-2008. The linkage output would be, for each patenting enterprise, a pair composed by the 'Applicant Identification Number in PATSTAT' and the 'Enterprise Identification Number in ASIA'. This last allows for accessing the repositories of the official statistical data and, therefore, linking Istat economic data to patenting enterprises. For example, factors influencing patenting propensity of enterprises might be studied, as well as the economic impact of patenting activity.

On the methodological side, linkage of patent data has to rely on the 'applicants names'. Consequently, a great effort has been put in the pre-processing phase to standardise the applicant/enterprise names and extract the 'legal form' from the name string. During the linkage process, two practical problems were faced: the reduced number of comparison variables and the huge dimension, in terms of number records, of the ASIA. These issues were addressed within a rule-based deterministic record linkage approach. In this paper, we will illustrate the main features of the adopted sequential searching and linkage methodology.

The paper is organised in 4 sections. In section 1 a description of ASIA and PATSTAT databases is provided. In section 2, details on the record linkage methodology as applied to these particular datasets are reported. The emphasis is put on search space reduction methods due to the number of comparison variables and to the huge amount of data. In section 3, some preliminary results are shown. In the last section, some conclusions and ideas for further improvements are given.

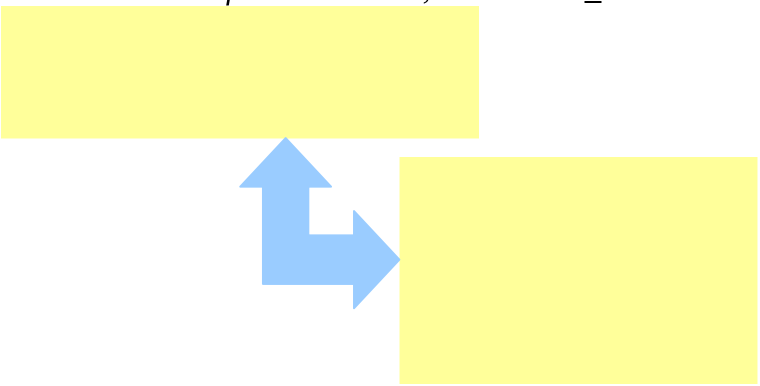
2. Registers: Italian business registers and patenting persons

A patent is an exclusive right granted by an authorized patent office for an invention, which is a product or a process providing a new (technical) solution to a problem. A patent provides protection for the invention to the owner of the patent. The first step in securing a patent is the filing of a patent application that involves three main actors: the inventor, the owner and the applicant.

The EPO database "Worldwide Patent Statistical Database", called PATSTAT, is probably the most complete and updated database on patents and patents applications. PATSTAT is updated twice a year, contains 20 tables organized as a relational database with more than 70 millions of records from over 80 countries. In this work only the two tables depicted in Figure 1 are considered. The link between them is given by the unique values of the field Applicant Identification Number, AIN. The AIN also contains the patent year of registration. The time period covered by the database is 1985-2010. PATSTAT registers both the inventor and applicant name; only the latter was used in this work. There is no explicit database field concerning the legal form of the inventor, owner or applicant. The possible legal form has been therefore extracted from those names. About the applicant, PATSTAT also registers its address (street, city, postal code) and its country code. Only applicants based in Italy, i.e. COUNTRY_CODE = "IT", were selected. In this work, the postal code was used as geographical location assuming it has the same accuracy as the address.

About the patent, PATSTAT registers its IPC (International Patent Classification), its application and publication number. It is worth noting that a patent could have assigned more than one IPC codes. It should also be stressed that there is no formal/well-defined relationship between IPC codes and the principal economic activity classification (NACE). Additional details on PATSTAT may be found at www.epo.org.

Figure 1: *Used database tables from PATSTAT; COUNTRY_CODE = "IT"*



Applicants may be individuals or establishments. The latter, according to the Frascati manual, see OECD (2002), could be: business enterprises, public institutions, non-profit institutions and private or public universities. In this work, the identification of patenting business enterprises is addressed.

On enterprises, the Istat business register ASIA, is considered. ASIA is developed, updated and maintained through the statistical integration of different administrative sources (Tax Register, Social Security Register, etc.), covering the entire population of enterprises of industry and services. Among the variables included in ASIA, one may specify:

a) *Enterprises Identification Number, EIN*, (an Istat internal and unique identification number allowing linkage to whatever economical information on the same unit collected by Istat);
b) *Enterprises Name*
c) *Zip Code*
d) *NACE code*
e) *Geographical information* (address, municipality, province, region),
f) *Legal form*

It has to be observed that ASIA and PATSTAT variables overlap only on Enterprise Name and Zip Code. Only enterprises being active in the period 1998-2008 have been analyzed (the size of ASIA varying from 3.8 to 4.5 millions of records). Considering that enterprises showing a high research and innovation propensity could have higher patenting propensity, a preliminary investigation has been conducted on the list frame of Research and Development survey (a subset of ASIA).

3. Development of a record linkage process

PATSTAT counts 299769 applications based in Italy and identified by an AIN. The number of non duplicated application numbers reduces to 72037. To each AIN in the PATSTAT database, an applicant name and the Zip Code are assigned. Additional variables may be derived from the previous information: year of application, year of first/last application by applicant; number of patent applications filed by each applicant, region of residence of the applicants, etc.

The variable *Applicant Name* has been subject to the following standardisation operations:

1. transformation of all letters in upper case letters
2. removal of punctuations: (accents, symbols and special characters, double spaces; dots)

3. standardisation of known abbreviations (e.g. we found about 150 ways to say “in short”)
4. standardisation of the most frequent words using equivalence lists, a deterministic record linkage procedure in Relais, see Istat (2011)
    a) input files: a file of words with frequencies greater than 1000; a file of words with frequencies greater than 100, but smaller than 1000;
    b) parameters: comparison function = “Edit distance”; threshold=0.8, greedy algorithm to perform the one-to-one assignment;
    c) output check: the word pairs declared “match” were subject to a clerical review;
    d) output: 122 matched pairs standardized; generally concerning singular/plural or Italian/English translation (for example: SERVICES/SERVIZI) ;
5. removal of duplicated words in the same name;
6. ordering of words in alphabetical order;
7. identification and standardisation of the legal form and storage in a variable called *Legal Form*. About 80 ways of expressing 6 main standardized legal forms were identified. In table 1, the distribution of the variable *Legal Form* is shown. Around 40% of the records have none legal form, while the majority (about 56%) is concentrated in “LTD” categories.

The same pre-processing was applied to ASIA. The resulting variable is called *Standardized Name* and has been used as comparison variable together with *Zip Code* and *Legal Form*.

The PATSTAT data file has been de-duplicated (by records having simultaneously the same values for the three comparison variables). Thus, the number of records reduced from 72037 to 23833.

Table 1: *Distribution of Legal Form, PATSTAT database*

Legal Form		**COOP**	**SAS**	**SNC**	**SPA**	**SRL**	**Total**
Frequency	8979	63	501	756	6164	7370	23833
%	37.67	0.26	2.10	3.17	25.86	30.92	100

The linkage output should be the pair *AIN* (PATSTAT) - *EIN* (ASIA). The latter allows linkage of structural and economical information stemming from Istat official surveys to patenting enterprises.

In PATSTAT, *Applicant Name* is missing in 40 records, while *Zip Code* is missing in about 10% of records. Besides the missing value problem, variable *Zip Code* in PATSTAT, also presents about 9.4% of values representing the geographical location only at aggregated level.

3.2 Search space reduction

Due to the size of ASIA, the amount of candidate matching pairs is huge and the usage of search space reduction techniques has been necessary. In this section details on the search space reduction techniques applied to PATSTAT and ASIA are given. Moreover, a blocking technique by neighbourhoods of words is introduced. Some classical blocking techniques based on the patent year or 2-digit *ZIP Code* were not effective; these are not further detailed here.

PATSTAT was reduced in order to contain only units probably representing enterprises. A list of Italian First Names, containing about 1600 records was used. From the PATSTAT database, we removed those records whose *Standardized Name* satisfy simultaneously the following conditions: a) it contains an Italian First Name b) it has an empty *Legal Form* and c) it does not contain special words indicating a business activity (e.g. enterprise, systems, etc.). About 63 such special words were found.

PATSTAT was then divided in two parts: 7700 records considered non-enterprises and 16132 records declared enterprises. The record linkage process was applied to the latter.

The eleven datasets of ASIA (1998-2008) were prepared in such a way that an active enterprise is included only once in their union. ASIA 2008 was the most complete and updated version.

In the union of different waves of ASIA, except for 885 records (over more than 7 millions), the *ZIP Code* is always registered with 5 digits.

Due to the huge computational burden, ASIA 2008 was divided in three parts: a) with more than 10 employees, b) with 1-9 employees, with non-empty *Legal Form*, and c) with less than 1 employee with non-empty *Legal Form*.

None of the comparison variables was considered reliable enough to be used as blocking variable. The idea of **neighbourhood of words** was then introduced. For a pair of records, it was assumed that a necessary matching condition is that their *Standardized Name*s share at least one exact word. Thus, it was assumed that at least one word is registered correctly. Then, for each PATSTAT record, the list of words forming its *Standardized Name* was found. Next, for each such a word, the list of enterprises in ASIA containing it was identified. The union of lists of such enterprises was named *Neighbourhood* of the *Standardized Name* under consideration. If an exact match on *Standardized Name* exists, it should belong to this *Neighbourhood*. For each record in PATSTAT, the record linkage procedure was applied using the *Neighbourhood* as blocking variable.

Blocking by *Neighbourhood* allows us to divide the search space in a huge number of much smaller search spaces. Obviously, the number of search spaces equals the number of records in PATSTAT and RELAIS may deal with many search spaces in an automatic manner. Each search space has a reduced size. The maximum size of such search spaces equals 15570, a very reasonable size to deal with in record linkage problems. By construction, each *Neighbourhood* contains at maximum one correct link. Due to this reason and to the dependency between *Neighbourhood* and *Standardized Name* variables, this blocking procedure, as it was defined here, could hardly be used in any probabilistic record linkage (missed independence).

Names having the longest word less than 2 characters were excluded from the search space creation as they could create huge *Neighbourhood*s (as very common words can also do). Moreover, it might happen that some *Standardized Names* have an empty *Neighbourhood*. This is generally the case for *Standardized Names* of a single word (if such words are differently registered in PATSTAT and ASIA for example). Of course, neighbourhoods could be defined also by an approximate matching (e.g. a similarity distance different instead of equality distance) of at least one word.

3.3 Deterministic record linkage

Even if the *Neighbourhood* was used as blocking variable, a similarity criteria between *Standardized Names* was used to give an overall measure of the records similarity. A compound deterministic rule was used: at least one of the following string comparators should be greater than 0.8: Jaro; Levensthein; Jaro-Winkler; Dice; 3-Grams; equality rule[1]. The selection of the unique links was performed using a greedy solution

[1] Details on the implementation of this comparison functions may be found in the RELAIS manual.

implemented in RELAIS. Equal weights for all rules were always used. Finally, the pairs declared matches were subject to a clerical review.

4. Results and exploration of possible analysis

At this stage, the number of found “correct” link is 13526 out of 16132, i.e. 84%. As for “correct” link we intend a (non duplicated) pair *AIN* - *EIN*; the pairs declared links have been clerically classified as “correct”, “possible links” (possibly subject to more detailed and sophisticated clerical review) or “false” (discarded). Even if pairs are non-duplicated, some of them may represent duplication of *Applicants* (more than one *applicant* may be linked to the same *enterprise*. This situation might happen when a multi patenting applicant has been registered with different names in different applications and the standardisation process does not compensate for these differences.
To asses the quality of the results, a short experiment has been conducted on a set of 190 codes randomly selected from the Espacenet web database (the *AIN* field has been used to download patent information from the EPA web-site[2]). We found 5 mismatches out of 190 records (2,5%). This means that, even if the available standardised information coincide in the two sources it is not possible to guarantee 100% exact link because of very similar (or common) names. Other possible sources of misclassification that should be taken into account when checking the quality of the linkage process are: enterprises belonging to the same enterprise group often register their patents with similar names; the changes occurred to enterprises through their life (changes of address, legal form, etc.).

Table 2: *Distribution of patenting enterprises which are active in 2009.*

	NACE	SizeClass				Total	
		1	[2-9]	[10-99]	[100-		%
1	**28**	77	255	1121	427	1880	20.6
2	**25**	38	130	493	142	803	8.8
3	**46**	155	292	287	34	768	8.4
4	**22**	23	76	327	124	550	6.0
…	**...**	…	…	…	…	…	
25						103	
73							
	Total	1251	1866	4202	1786	9105	
	%	13.74	20.49	46.15	19.62		
	% Pop	58.44	36.49	4.81	0.26		

The structure of the patenting enterprises is the first joint analysis that could be performed once the linkage between patents and enterprises is found. In Table 2 the patenting enterprises being active in 2009 are reported by size class and Nace code. The 9105 patenting enterprises still active in 2009 are distributed over 73 NACE divisions. Only 25 of these divisions show a frequency greater that 100. Only the most frequent

[2] Ten applicants number can be downloaded by trials, for a maximum of 200 AINs. Moreover, the information provided need to be managed before the use.

NACE divisions are shown in table 2. We observe (the last column of the table) that about 44% of the patenting enterprises are concentrated in Nace 28, 25, 46 and 22,
i.e. Manufacture of machinery and equipment n.e.c., Manufacture of fabricated metal products, except machinery and equipment, Wholesale trade, except of motor vehicles and motorcycles and Manufacture of rubber and plastic products, respectively. In the last two rows of table 2, we observe that, as expected, more than half of the population of patenting enterprises (65%) have a size greater or equal to ten employees (the highest class considered), while the population of enterprises with more than 10 employees represents only 5% of the entire population of enterprises.

The second type of analysis is represented by the study of some special subpopulations. As an example, one could analyse the structure of enterprises patenting in the biotech domain. This analysis is possible since information on the structure of enterprises (principal economic activity and/or number of employees) is available in the ISTAT business registers while information on the biotech-related patents may be retrieved from the PATSAT database. 204 enterprises among the 9105 enterprises being active in 2009 applied for a biotech-related patent. These 204 enterprises are distributed over 28 Nace divisions, but two of them cover by themselves more than a half of the biotech-patenting enterprises. These two Nace divisions are 21 and 72, i.e. Manufacture of pharmaceuticals, medicinal, chemical and botanical products and Scientific research and development, respectively. The distribution of biotech-patentitng enterprises is shown in table 3.

Table 3: *Distribution of biotech-patenting enterprises which are active in 2009.*

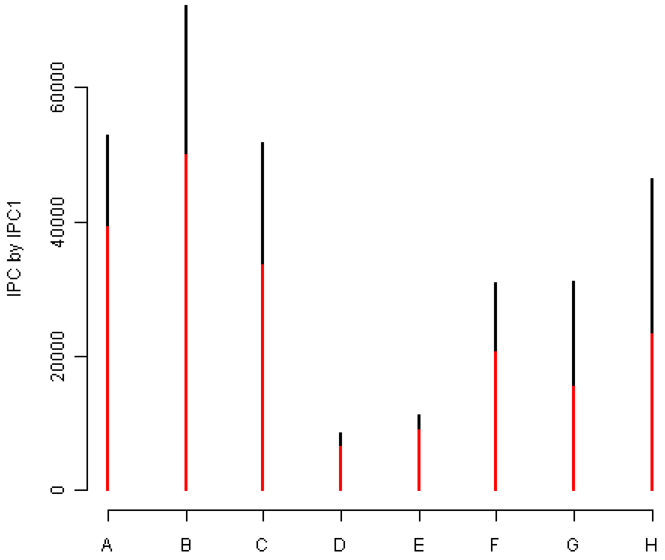
	Nace	SizeClass				Total
		1	[2-9]	[10-99]	[100-	
1	**21**	2	1	14	36	53
2	**72**	9	20	18	3	50
3	**20**	0	6	8	7	21
4	**46**	0	4	11	4	19
28						
	Total	20	44	71	69	204
		9.8	21.6	34.8	33.8	

As previously detailed, the PATSTAT database contains about 300.000 record related to the patents obtained by Italian applications. Once the link between enterprises and patents applications is established, it is possible to observe that about 90% of applications are performed by enterprises. The patents classification according to the IPC code is not related to the Nace classification. The classification according to the IPC code follows a hierarchical structure which is described at www.epo.org. The IPC-letter distribution of the 300.000 applications of Italian enterprises is shown in table 4. In figure 2, the distribution of the IPC codes that were found by the record linkage process is shown in red, while the original distribution of the same IPC codes is shown in black.

Table 4: *Distribution of the IPC of the Italian applications of enterprises.*

IPC	A	B	C	D	E	F	G	H
	HUMAN NECESSITIES	PERFORMING OPERATIONSTRANSPORTING	CHEMISTRY; METALLURGY	TEXTILES; PAPER	FIXED CONSTRUCTIONS	MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING	PHYSICS	ELECTRICITY
% Pop	17	24	17	3	4	10	10	15

Figure 2: *Distribution of original IPC codes (black) and linked IPC codes (red)*



5. Conclusions and future plans

In this paper we illustrated the path followed at Istat in designing a linkage strategy to match micro-data on patent applications from PATSTAT and ASIA. The overall aim of this project is to identify the Italian patenting enterprises and characterise them through their economical information.

In PATSTAT, the applicants resident in Italy and registering at least one patent in the period 1985-2010 have been considered. Patent applicants can be 'individuals' or 'establishments'. At this stage, the linkage process aimed at identifying, among establishments, the business enterprises recorded in ASIA in the period 1989-2009.

The overlapping information between the two archives reliable as matching variables in the linkage process mainly consists only of the 'applicants names' and the 'postal code'. Moreover, the size of the business register ASIA in terms of number of records represents a computational problem to be faced. Therefore, a great effort has been put in the pre-processing phase of the process to standardise the applicant/enterprise names and some 'search space' reduction techniques have been adopted. Among the latter, particularly effective has proved to be the ‘blocking by neighbourhood’ technique. Assuming that, for a given patenting enterprise, at least one word in the ‘applicant name’ (in PATSTAT) and the ‘enterprise name’ (in ASIA) is correctly registered in both the archives, the ‘neighbourhood’ of an applicant name is defined as the set of enterprises which have a name containing at least one word equal to a word in the applicant name. Then, the correct link for the given applicant have been searched within its neighbourhood.

At this development stage, around 84% of patenting enterprises (13526 out of 16132 applicants) were identified as ‘establishments’. The next step will be to define the ‘neighbourhood’ on the base of similarity between words instead of equality, in order to manage typing errors. Some further improvements might be obtained by using the address instead of the *Zip Code*.

In future work, it would be desirable to classify the whole set of patenting establishments as: business enterprises, public institutions, non-profit institutions and private or public universities enterprises, according to the Frascati Manual (2002). For applicants without legal form, it is planned to use different archives (such as the List of enterprise manager or the List of companies partners). Finally, a probabilistic approach to the record linkage could be derived by using the R&D survey frame test set.

References

OECD (2002). *Frascati Manual 2002: Proposed Standard Practice for Surveys on Research and Experimental Development*, Paris 2002.

Istat (2003), *Metodi statistici per il record linkage*, Metodi e Norme n. 16, Anno 2003, a cura di Mauro Scanu

Istat (2011), *RELAIS - Record linkage at Istat*, software and User’s guide available at: http://www.istat.it/strumenti/metodi/software/analisi_dati/relais/