

**Creating Powerful Indicators for Innovation Studies with Approximate Matching Algorithms. A test based on PATSTAT and Amadeus databases**

Grid Thoma

Department of Political Science and Law Studies, University of Camerino, Italy  
and

CESPRI, Bocconi University, via Sarfatti 25 20136 Milano  
grid.thoma@unibocconi.it

Salvatore Torrisi

Department of Management, University of Bologna  
via Capo di Lucca 34 40126 Bologna, Italy

and

CESPRI, Bocconi University, via Sarfatti 25 20136, Milano  
torrisi@unibo.it

PRELIMINARY DRAFT

September 2007

please do not quote without authors' permission

paper prepared for the *Conference on Patent Statistics for Policy Decision Making*  
2-3 October 2007 San Servolo, Venice

## Abstract

The lack of firm-level data on innovative activities has always constrained the development of empirical studies on innovation. More recently, the availability of large datasets on indicators, such as R&D expenditures and patents, has relaxed these constraints and spurred the growth of a new wave of research. However, measuring innovation still remains a difficult task for reasons linked to the quality of available indicators and the difficulty of integrating innovation indicators to other firm-level data.

As regards quality, data on R&D expenditures represent a measure of input but do not tell much about the ‘success’ of innovative activities. Moreover, especially in the case of European firms, data on R&D expenditures are often missing because reporting these expenditures is not required by accounting and fiscal regulations in some countries.

An increasing number of studies have used patents counts as a measure of inventive output. However, crude patent counts are a biased indicator of inventive output because they do not account for differences in the value of patented inventions. This is the reason why innovation scholars have introduced various patent-related indicators as a measure of the ‘quality’ of the inventive output. Integrating these measures of inventive activity with other firm-level information, such as accounting and financial data, is another challenging task. A major problem in this field is represented by the difficulty of harmonizing information from different data sources. This is a relevant issue since inaccuracy in data merging and integration leads to measurement errors and biased results.

An important source of measurement error arises from inaccuracies in matching data on innovators across different datasets. This study reports on a test of company names standardization and matching. Our test is based on two data sources: the PATSTAT patent database and the Amadeus accounting and financial dataset. Earlier studies have mostly relied on manual, ad-hoc methods. More recently scholars have started experimenting with automatic matching techniques. This paper contributes to this body of research by comparing two different approaches – the character-to-character match of standardized company names (perfect matching) and the approximate matching based on string similarity functions. Our results show that approximate matching yields substantial gains over perfect matching, in terms of frequency of positive matches, with a limited loss of precision – i.e., low rates of false matches and false negatives. Finally, we find that taking into account the priority links between USPTO patents and EPO patents yields a significant gain in the number of EPO matched applications.

**Acknowledgments.** We thank Jim Bessen, Rachel Griffith, Dominique Guellec, Bronwyn Hall, Dietmar Harhoff, Gareth Macartney, Tom Magerman, Bart Van Looy, Bob Reijna, James Rollinson, Colin Webb, Maria Pluvia Zuniga, and all the participants at the PATSTAT Users’ Meeting in Geneva in June 2007 for very fruitful discussions during the preparation of this paper. We also thank Armando Benincasa and Luisa Quarta from Bureau Van Dijk for clarifications about the structure of the Amadeus database and its changes over time. Data collection and elaboration reported in this work was partly carried out during the ongoing European Commission project “Study of the effects of allowing patent claims for computer-implemented inventions”. The opinions expressed in this publication are those of the authors and do not necessarily reflect in any way opinions of the European Commission or any of the partners.

## 1. Introduction

Until recently empirical studies on the economics and management of innovation have suffered from a paucity of data at the firm level.

Scholars of technical change have addressed the lack of data by following two directions. A first approach has tried to collect firm-level information through surveys based on representative samples of the population of innovators.

Regarding the US context two widely cited surveys are the Yale survey (Levin et al 1987) administrated in the early 1980s and its subsequent version conducted by scholars at the Carnegie Mellon University in the 1990s (Cohen et al 2000) . These two surveys provide an useful source of detailed information on the nature and strategies of innovation and the means used to appropriate the economic returns generated innovative activities.

Similarly, in the European context the Community Innovation Survey (CIS) collects detailed data on innovation and other firm characteristics such sales, employment, exports/imports, etc. Unlike Yale and Carnegie Mellon surveys, which have been administrated by academic researchers, the CIS is conducted by National Statistical Offices with the aim of achieving a large coverage of industries and types of innovators (large and small firms etc.) (Arundel, 2003). Unfortunately, integration of CIS data with other information, like patents and accounting data is made difficult by the limitations to the use of CIS data imposed by privacy laws in countries like Italy. These shortcomings of the CIS dataset limit its use for the purposes of research in economics, management and public policy. More recently, scholars have conducted new innovation surveys providing very detailed information on the factors driving innovation at the level of individual inventors (Harhoff et al. 1999; Gambardella et al., 2000; Giuri, Mariani et al., 2007).

Another research line has focused on the collection of information on different qualitative dimensions of innovation such as *prizes* as a measure of successful inventive races, *trademarks* as a measure of the new product introduction, newswires as a paper trail of patterns of collaborations among firms such as M&A, licensing and R&D agreements etc. (Moser, 2004; Giarratana and Torrissi, 2006; Fosfuri and Giarratana, 2007; Powell et al., 2000; Arora, Fosfuri, Gambardella, 2001) A third line of exploration is centred on innovation counts and R&D. R&D expenditures are a measure of input and do not tell much about the ‘success’ of innovative activities. Moreover, especially in the case of European firms, data on R&D expenditures are often missing because reporting these expenditures is not required by accounting and fiscal regulations in some countries. An increasing number of studies have used patent counts and patent-related indicators to measure the quantity and the ‘quality’ of inventive output. Patents as a measure of inventive success have their own drawbacks too but they are the most direct and objective measure of innovation (Griliches, 1981 and 1990; Pavitt, 1988).

Patent analysis has been pioneered by Zvi Griliches and colleagues (Griliches, 1981 and 1990; Griliches, Hall and Pakes, 1991) at the National Bureau of Economic Research (NBER) and by Keith Pavitt and colleagues at the Science Policy Research Unit (SPRU) -University of Sussex (Pavitt, 1985 and 1988; Patel and Pavitt, 1991). The NBER patent dataset on US data has represented a path-breaking effort in this field providing new data that are useful to account for differences in the ‘value’ of patents (Hall, Jaffe and Trajtenberg (2001 and 2005). Bronwyn Hall and colleagues have made public the NBER patent citation database. They have also disclosed to the research community the links between the names of USPTO patent assignees with the names of US companies listed in the Compustat dataset.

A major obstacle to the integration of patent data with other indicators of firm performance in large samples is represented by the difficulty of univocally matching the names of patent assignees with the corresponding legal entity in business directories such as Compustat or Who Owns Whom. Previous studies have addressed this issue by trying automatic matching procedures to reduce the cost of data standardization and integration.

The first step in this setting is represented by name standardization. To our knowledge, the most important attempts at standardizing patentee names are the Thomson Scientific's Derwent World Patent Index (2002) and the USPTO's CONAME standardization files. More recently, another standardization method has been developed by a group of researchers from the K.U. Leuven for the Eurostat (Magerman, Van Looy and Song, 2006).

The Derwent Index is constructed by assigning a code to 21,000 patentees. This index accounts for legal links between parent companies and subsidiaries thus achieving a legal entity standardization. This is made possible by the use of information on corporate structure collected from secondary business sources. This includes also information on M&As, changes of names and reorganization (e.g., new subsidiaries). Legal entity standardization requires substantial manual, labor-intensive work and some loss of accuracy in name matching thus giving rise to a potentially large number of false positives. Moreover, the process leading to standard names and in case of M&As and name changes the criteria adopted for name standardization are case specific (Magerman, Van Looy and Song, 2006).

The CONAME file compiled by the US Patents and Trademarks Office is a semi-automatic standardization procedure which focuses on the first-named assignee reported in the patent document. For patents granted after July 1992 the assignee name is standardized and matched automatically with other standardized names in the same dataset. New assignees that are not matched automatically with standardized names in the dataset are matched manually. For instance, the entry of a new assignee whose standardized name does not match any previously standardized names is examined by looking at the names of inventors. The CONAME file accounts for changes in assignee names but does not account for legal links between assignee names. Moreover, similar names with a different legal form or from different countries are not matched.

The K.U.Leuven (KUL) methodology consists in the standardization of patentee names and perfect matching of names. The advantage of this method is a high level of accuracy at the cost of some loss of completeness. This is a conservative, fully automatic methodology which, like the CONAME file, does not try to establish links between similar names neither it seeks to find legal links among assignees.<sup>1</sup> The main advantage of this procedure is high precision, i.e., a limited number of false matches. Inevitably, this method does not fare well in terms of completeness since a high number of good matches may remain unmatched. The KUL methodology has been used to standardize and match assignee names of EPO patent applications published between 1978 and 2004 and USPTO granted patents published between 1992 and 2003 (Magerman, Van Looy and Song, 2006).

Drawing on the Derwent methodology, Rachel Griffith, Gareth Macartney and colleagues at the Institute of Fiscal Studies (IFS) have standardized the names of a sample of UK assignees of Triadic patents and matched them with the standardized names of companies contained in Bureau van Dijk's Amadeus database. Only identical standardized names found in the two datasets are matched by the IFS using the Derwent semi-manual standardization procedure.

We have conducted a matching test by comparing assignee names in the PATSTAT dataset with company names in the Amadeus dataset for a sample of around 2,197 European publicly listed firms and their 146,728 subsidiaries. These firms have disclosed information on their R&D expenditures. Comparing these data with the OECD R&D STAN database we found that these companies account for around 90% of the total intramural business R&D expenditures in the European countries in year 2000.

The names found in the two datasets are standardized using a variant of the KUL methodology and then matched by the Jaccard similarity string function (Jaccard 1901).<sup>2</sup> Our experiment shows that

---

<sup>1</sup> The term standardization here is used to refer to all operations required to produce a list of standardized names like the Derwent standard codes. Harmonization is used to mean the integration of (standardized or non-standardized) names from different datasets to obtain codes which uniquely identify given legal entities (e.g., Fiat S.p.a. and its subsidiaries COMAU and CNH).

<sup>2</sup> The matching program in Java was developed by a colleague at the Computer Science Dept of Bologna University.

approximate string matching (ASM) yields a substantial gain over perfect matching in terms of number of patent assignees found in the Amadeus dataset. However, these gains are obtained at the cost of a loss of accuracy. Depending on the level of precision which one aims to achieve, matching similar names implies a higher risk of false matches as compared with perfect matching. We estimated the number of false positives and false negatives at different levels of the Jaccard similarity (J) score by manually inspecting all matched names corresponding to different levels of the J distance. To estimate the incidence of false positives we checked all occurrences for levels of the J distance above 0.7 and found that the maximum number of false positives represents less than 6% of total matches. The motivations for choosing 0.7 as a threshold are explained in the paper. To estimate the incidence of false negatives we looked at EPO assignees with more than 15 patents and found that 8.5% of these have not been matched by using the Jaccard measure. These results suggest that using the approximate matching methodology yields significant improvements in terms of completeness at the price of a relatively small cost in terms of loss of precision.

The paper is organized as follows. Section 2 describes the dataset while Section 3 illustrates the methodology. The results of the matching experiment are reported in Section 4 while Section 5 focuses on the results of some robustness checks. Section 5 analyzes the advantages from linking the USPTO patent assignees dataset with the EPO applicants dataset. Section 6 concludes.

## 2. Data

Our analysis is based on the links between two datasets. The first data source is Bureau Van Dijk's Amadeus, a dataset containing accounting and financial information of about 9 million firms from 34 EU and Eastern European countries. For each firm longitudinal data are available for a period of up to ten years. Amadeus draws its information from about 50 country providers, which in most cases are the national registers of companies.<sup>3</sup>

The main advantage of Amadeus over other data sources is its coverage of small and medium sized firms for a large set of countries.

Company data are harmonized by an identification number (the BVD number), which allows to identify uniquely a given business legal entity. The BVD number is based on standard national codes such as the registry number or VAT firm number. A BVD number is also available for most subsidiaries of company groups. In the case of groups Amadeus provides information on ownership links between parent companies and subsidiaries. In most European countries, publicly listed firms and corporations with consolidated accounts should report the complete list of subsidiaries - i.e., those firms that are controlled *de jure* (51% of shares) or *de facto* (the parent company directly or indirectly owns a share of the firm's assets that guarantees an effective control).

The links between parent companies and subsidiaries are the main source used by BVD for constructing corporate structure. Moreover, changes in ownership structure due to mergers, acquisitions or spin-offs are taken into account by BVD. Detailed information on these changes is reported in the Zephyr dataset, another BVD dataset containing a stock of about 400,000 worldwide deals in 2007.

For publicly-listed firms, BvD collects directly around 20 thousands annual reports worldwide (BvD, 2006).

For our purposes we used the Amadeus dataset for the period 1997-2005. Before 1996 information on corporate structure reported in Amadeus is less complete and reliable.<sup>4</sup>

Our source of patent data is the *EPO Worldwide Patent Statistical Database* (PATSTAT), which is available under license from OECD-EPO Task Force on Patent Statistics. PATSTAT not only

---

<sup>3</sup> The list of the national providers is available at [www.bvdep.com](http://www.bvdep.com).

<sup>4</sup> From a conversation with the Italian subsidiary of BVD in Milan we understood that Amadeus has become a commercial product in 1996.

includes data on patent indicators such as citations and IPCs codes, but also on patent families based on priority links.

Our matching exercise is centered on 2,197 European publicly-listed firms which have disclosed information on their R&D expenditures. R&D data were collected from various sources, including BVD's Amadeus, Compustat's Global Vantage and the UK Department of Industry's R&D Scoreboard.

Amadeus made it possible to track all changes in names and corporate structure over the period 1997-2005. After these checks we ended up with around 146,728 distinct subsidiaries. For 130 firms out of 2,197 we could not find any subsidiary. Table 1 reports the sectoral distribution of parent companies, their subsidiaries by the sector of the parent, and the relative amount of R&D expenditures. The total number of subsidiaries in Table 1 is larger than 146,728 because of double counting. In particular we found that 5251 subsidiaries – around 3,5% - are controlled by more than one parent company.

As Table 1 clearly shows, the sample of firms is concentrated in few sectors such as software, electronic instruments and telecommunications equipment, computers, electrical machinery, chemicals and pharmaceuticals. The distribution of subsidiaries is still quite concentrated but in different sector like public utilities, food and tobacco and motor vehicals and telecommunication services. Moreover, over 75 per cent of R&D expenditures are accounted for five sectors. Overall, the sample firms are representative of the most R&D-intensive sectors in Europe.

It is important to notice that the sample firms account for about 87 per cent of total business R&D in the top 25 European countries (see Table 2).

**Table 1 Distribution of Firms, Subsidiaries and consolidated R&D expenditures**

2,5 digit industry class	Firms with R&D		Subsidiaries		R&D expenditures	
	N	%	N	%	Mil EUR	%
01 Food & tobacco	87	3,96	11784	7,75	36213	3,4
02 Textiles, apparel & footwear	45	2,05	2025	1,33	1621	0,1
03 Lumber & wood products	10	0,46	590	0,39	76	0
04 Furniture	21	0,96	1103	0,73	3746	0,4
05 Paper & paper products	30	1,37	3105	2,04	3671	0,3
06 Printing & publishing	27	1,23	2513	1,65	1448	0,1
07 Chemical products	92	4,19	10296	6,77	113508	9,8
08 Petroleum refining & prods	38	1,73	5158	3,39	42962	3,1
09 Plastics & rubber prods	38	1,73	2679	1,76	11301	0,9
10 Stone, clay & glass	47	2,14	7669	5,05	8584	0,7
11 Primary metal products	55	2,5	5558	3,66	11171	0,6
12 Fabricated metal products	60	2,73	3974	2,61	5363	0,3
13 Machinery & engines	171	7,78	7935	5,22	37780	2,3
14 Computers & comp, equip,	50	2,28	1734	1,14	5289	0,4
15 Electrical machinery	78	3,55	7601	5,00	138838	13,4
16 Electronic inst, & comm, eq,	224	10,2	7258	4,78	161218	13,1
17 Transportation equipment	18	0,82	2547	1,68	46167	4
18 Motor vehicles	53	2,41	8396	5,52	257633	19,6
19 Optical & medical instruments	75	3,41	2565	1,69	11981	0,9
20 Pharmaceuticals	131	5,96	5309	3,49	258664	18,3
21 Misc, manufacturing	37	1,68	1371	0,90	2369	0,2
22 Soap & toiletries	17	0,77	2586	1,70	12480	1,2
24 Computing software	326	14,84	7934	5,22	31276	1,3
25 Telecommunications	48	2,18	6464	4,25	29817	2,5
26 Wholesale trade	53	2,41	1856	1,22	1712	0,1
27 Business services	50	2,28	1798	1,18	4860	0,4
28 Agriculture	3	0,14	34	0,02	1	0
29 Mining	29	1,32	2275	1,50	4017	0,2
30 Construction	42	1,91	5433	3,57	8357	0,4
31 Transportation services	17	0,77	3336	2,20	4817	0,4
32 Utilities	58	2,64	12436	8,18	37931	1,1
33 Trade	23	1,05	1324	0,87	472	0
34 Fire, Insurance, Real Estate	27	1,23	1095	0,72	935	0
35 Health services	9	0,41	124	0,08	454	0
36 Engineering services	85	3,87	2623	1,73	5275	0,3
37 Other services	23	1,05	1491	0,98	563	0
Overall	2197	100	151979	100,00	1302570	100

**Table 2. Distribution of R&D expenditures by country and by sector**

Country	Year	R&D expenditure in millions of euros					As a share of total expenditure				Our sample relative to		
		Business Sector	Govt Sector	HEI Sector	Other	Total R&D	Our sample	Business Sector	Govt Sector	HEI Sector	Other	Business sector	Total R&D
Austria	2002	3131	266	1266	21	4684	676	66,8%	5,7%	27,0%	0,4%	21,6%	14,4%
Belgium	2000	3589	312	1005	58	4964	940	72,3%	6,3%	20,2%	1,2%	26,2%	18,9%
Bulgaria	2000	15	49	7	0	71	0	21,4%	68,6%	9,8%	0,2%	0,0%	0,0%
Switzerland	2000	5065	90	1566	132	6852	10086,16	73,9%	1,3%	22,9%	1,9%	199,2%	147,2%
Cyprus	2000	5	11	6	2	25	0	21,3%	46,6%	24,8%	7,3%	0,0%	0,0%
Czech Rep.	2000	446	188	106	4	744	1,098279	60,0%	25,3%	14,2%	0,5%	0,2%	0,1%
Germany	2000	35600	6873	8146	0	50619	34871,29	70,3%	13,6%	16,1%	0,0%	98,0%	68,9%
Denmark	2000	2596	492	770	34	3892	1227	66,7%	12,6%	19,8%	0,9%	47,3%	31,5%
Estonia	2000	8	9	19	1	37	1	22,5%	23,1%	52,4%	1,9%	11,4%	2,6%
Spain	2000	3069	905	1694	51	5719	21	53,7%	15,8%	29,6%	0,9%	0,7%	0,4%
Finland	2000	3136	468	789	30	4423	3690	70,9%	10,6%	17,8%	0,7%	117,7%	83,4%
France	2000	19348	5361	5804	439	30954	19258	62,5%	17,3%	18,8%	1,4%	99,5%	62,2%
Greece	2001	278	188	383	3	852	111	32,7%	22,1%	44,9%	0,4%	39,8%	13,0%
Croatia	2002	115	60	95	0	271	51	42,7%	22,2%	35,1%	0,0%	43,8%	18,7%
Hungary	2000	180	106	97	23	405	35	44,3%	26,1%	24,0%	5,6%	19,7%	8,7%
Ireland	2000	842	96	238	0	1176	517	71,6%	8,1%	20,2%	0,0%	61,4%	44,0%
Iceland	2000	142	64	41	5	251	4	56,4%	25,5%	16,2%	1,9%	2,7%	1,5%
Italy	2000	6239	2356	3865	0	12460	37	50,1%	18,9%	31,0%	0,0%	0,6%	0,3%
Lithuania	2000	16	31	27	0	73	0	21,5%	41,9%	36,5%	0,0%	0,0%	0,0%
Luxembourg	2000	337	26	1	0	364	2	92,6%	7,1%	0,2%	0,0%	0,5%	0,5%
Latvia	2000	15	8	14	0	38	0	40,3%	22,1%	37,6%	0,0%	0,0%	0,0%
Malta	2002	3	2	7	0	12	0	24,7%	16,4%	58,8%	0,1%	0,0%	0,0%
Netherlands	2000	4458	974	2120	75	7626	8582	58,5%	12,8%	27,8%	1,0%	192,5%	112,5%
Norway	2001	1814	444	780	0	3037	592	59,7%	14,6%	25,7%	0,0%	32,6%	19,5%
Poland	2000	432	386	377	2	1197	10	36,1%	32,2%	31,5%	0,1%	2,3%	0,8%
Portugal	2000	258	222	348	100	927	0	27,8%	23,9%	37,5%	10,8%	0,0%	0,0%
Romania	2000	103	28	18	0	149	0	69,4%	18,8%	11,8%	0,0%	0,0%	0,0%
Russia	2000	2087	721	134	7	2948	217	70,8%	24,4%	4,5%	0,2%	10,4%	7,4%
Sweden	2001	8118	297	2085	10	10511	7470	77,2%	2,8%	19,8%	0,1%	92,0%	71,1%
Slovenia	2000	167	77	49	3	297	34	56,3%	25,9%	16,6%	1,2%	20,4%	11,5%
Slovakia	2000	94	35	14	0	143	5	65,8%	24,7%	9,5%	0,0%	4,8%	3,2%
Turkey	2000	464	86	839	0	1389	0	33,4%	6,2%	60,4%	0,0%	0,0%	0,0%
UK	2000	18884	3672	5985	529	29070	19224	65,0%	12,6%	20,6%	1,8%	101,8%	66,1%
Europe	2000	121054	24902	38694	1528	186177	107661	65,0%	13,4%	20,8%	0,8%	88,9%	57,8%
EU15	2000	109883	22508	34499	1351	168238	96625	65,3%	13,4%	20,5%	0,8%	87,9%	57,4%
EU25	2000	111365	23436	35233	1385	171417	96711	65,0%	13,7%	20,6%	0,8%	86,8%	56,4%

US	2000	216552	29926	33221	10218	289917	0	74,7%	10,3%	11,5%	3,5%	0,0%	0,0%
Japan	2000	109181	15217	22354	7108	153860	0	71,0%	9,9%	14,5%	4,6%	0,0%	0,0%

Source: Eurostat and OECD (2007)

### 3. Method

Integration of patent data and accounting data consists of two main steps: name standardization and string matching.

Matching string fields usually involves two main steps: standardization and the actual matching phase. In the first step company names may require some preliminary cleaning before name standardization takes place. Names standardization requires a series of tasks like punctuation standardization (e.g., from FERRARI\_& C. to FERRARI\_& C.) and company name standardization (from FERRARI, & C. to FERRARI, AND COMPANY) (see Magerman, Van Looy and Song, 2006). String matching can be carried out by two different approaches: (a) character-to-character comparison; (b) more complex approximate string comparison techniques, which may increase the number of matches at the cost of a lower precision. It is worth to recall that a string is an ordered sequence of symbols or characters. In our case a string is a sequence of letters and characters that composes a company name.

#### *Data Preparation and Analysis*

As mentioned before, our analysis draws on two distinct sources of data: (a) a text file containing company names, company IDs (BVD numbers), parent IDs and countries names obtained from the Amadeus database for different years; and (b) and a file with patent assignee names and countries provided by the PATSTAT database.

Before starting name standardization and matching, the input files have been checked to correct for any character encoding, normalize the format (to make sure that data are in correct and comparable formats) and remove redundancies. These corrections are important to guarantee a proper application of the matching algorithms.

After this preliminary data cleaning stage we executed a manual inspection of a sample data to better understand the characteristic of the dataset and to find specific recurring names like COMPANY, LTD, &C., and CO. We also analyzed automatically the data to find punctuation symbols (e.g., ! “ & @ / and []), special text characters (e.g., Æ Ç È Ë Ä) and non-text characters, and an evaluation of string comparison methods on the specific data set. These preliminary tests serve the function of calibrating the standardization and matching operations.

Data analysis is also important to decide the most appropriate string similarity function(s) that should be used to match the names. String similarity functions compare two strings and produce a number ranging from 0 (= minimum similarity or maximum distance) to 1 (= maximum similarity or perfect matching). Among the various similarity functions, there are two that are worth to mention for their widespread use in the literature on data integration or harmonization (Navarro, 2001).

The first category of similarity functions is based on edit distance. For instance, the Levenshtein distance between two strings is defined as the minimum number of operations needed to transform a string into another one. The transformation of string can be obtained by character inserting, substituting, swapping or substitution (Levenshtein, 1966). An extension of the Levenshtein edit distance was developed by Smith and Waterman (1981). The main difference with the Levenshtein distance is that character mismatches at the beginning and the end of strings are ignored in the calculation of distance. For instance, two companies ‘Dr Michal White Plc’ and ‘Michael White Plc, Dr’ has a short distance using the Smith-Waterman distance.

The similarity between two strings  $x$  and  $y$  of length  $n_x$  and  $n_y$  can be calculated as  $1-d/N$ , where 1 is the maximum similarity,  $d$  is the distance between  $x$  and  $y$  and  $N=\max\{n_x, n_y\}$ . To calculate the distance between two strings we need to assign a cost  $c$  to each operation required to transform the string  $x$  into string  $y$  (or viceversa). The cost is 1 for substitution and deletion of a character and 0 for perfect matching characters. For instance, the edit distance between IBM and INTEL is  $1 - [c(I,I)+c(B,N)+c(M,T)+c(\emptyset,E)+C(\emptyset,L)]/5 = 1-4/5=1/5$ .

The second category of similarity functions rely on token-based distance. Measures of token distance, like the J similarity index, are based on the division of strings into tokens or sequences of characters. Token-based distance functions account for differences due to the position of the same tokens between otherwise identical strings (e.g., Peter Ross and Ross Peter).

To see which of these two similarity distance fit best our data we applied both measures to a small sample of data and analyzed manually the outcome of each matching procedure.

Using the edit distance, allowing substitution, deletion, insertion and character swapping, we found a series of problems that can be illustrated by using the following true examples:

1. HILLE & MUELLER GMBH & CO. /HILLE & MULLER GMBH & CO KG /HILLE & MÜLLER GMBH & CO KG
2. AB ELECTRONIK GMBH/AB Elektronik GmbH
3. BHLER AG /BAYER AG

The first two cases contain some *spelling variations* (e.g. Ü and UE) and *spelling errors* (k and C) respectively. While spelling variations can be approached by using edit distance functions with 0 transformations cost, spelling errors cannot be easily automatically identified without significantly reducing the precision of the method. However, these two case clearly show that the use of edit distances may increase the number of true positive matches compared with perfect match.

The third case illustrates an important drawback of this similarity function. The two strings have a low edit distance although they describe two unrelated companies. This demonstrates that an automatic application of edit distances to minimize the cost of string transformation (with only one or two operations) is made difficult by the distribution of company names in our dataset.

To test the performance of the second category of string similarity functions we used the J token distance after breaking the strings on white spaces and computing the fraction of common tokens.

$$J(x, y) = 1 - \frac{|x \cap y|}{|x \cup y|} = \frac{|x \cup y| - |x \cap y|}{|x \cup y|}$$

where  $|x \cap y|$  measures the number of common tokens between strings x and y while  $|x \cup y|$  measures the total number of distinct tokens.

Applying the J distance to our dataset yields the following potential matches:

1. AAE HOLDING /AAE TECHNOLOGY INTERNATIONAL
2. Japan as represented by the president of the university of Tokyo /President of Tokyo University
3. AAE HOLDING /AGRIPA HOLDING
4. VBH DEUTSCHLAND GMBH /IBM DEUTSCHLAND GMBH

The first two cases highlight the merits of similarity functions using the token-based distance. The third case shows that the database contains non-discriminating tokens like HOLDING which occur with a high frequency in our database. Non-discriminating tokens should be given a smaller weight than significant tokens like AAE in the matching process. Case 4 indicates that similarity functions centred on the token-based distance do not completely wipe out the problems found with similarity functions based on edit distance.

### *Name standardization*

The standardization procedure we adopted has been partially taken from Magerman, Van Looy and Song (2006). The main standardization operations can be divided into the following categories:

1. Character Cleaning
2. Punctuation Cleaning
3. Legal Form Indication Treatment
4. Spelling Variation Standardization

5. Umlaut Standardization
6. Common Company Name Removal<sup>5</sup>
7. Creation of an Unified List of Patentees

Unlike Magerman, Van Looy and Song (2006), who rely on a perfect matching approach, we did not remove white spaces in company names because these spaces are useful for calculating the token-based distance. Moreover we did not apply operations (6) and (7) because the use of the weighted J score allows us to overcome these steps. As we explain below, tokens with a high frequency in the dataset are assigned low weights and therefore have a small impact in the computation of the J Score. At the same time, maintaining common company names allows to fully use the information coming from PATSTAT and Amadeus and avoids the creation of a new ID index required in operation (7).

### *Matching*

As discussed before, character-to-character comparison of standardized strings yields a high level of precision at the cost of completeness. On the contrary, application of string distance functions may increase completeness at the cost of a lower precision.

To account for non-discriminating tokens we weighted each token proportionally to its frequency.

Formally, each token  $i$  is weighted with a weight  $w_i = \frac{1}{\log(n_i) + 1}$ , where  $n_i$  is the frequency of the

token in the dataset. This weighting method is a simplified version of the the tf-idf weight (term frequency-inverse document frequency) (Salton and Buckley, 1988).

Our similarity distance then is based on a modified J index that assigns to each token a weight inversely correlated with its frequency in the dataset. To reduce the computational complexity of the J similarity index we calculate it as follows:

$$\frac{2|x \cap y|}{|x| + |y|}$$

where the denominator is the sum of all tokens, including those tokens that are contained in both strings. This may result in some double counting. On other hand, it would be extremely costly from a computation viewpoint to find tokens common to two strings (company names). To correct in part for this problem we have multiplied the index by a factor of 2.

To illustrate the inverse relationship between the frequency of the token in the dataset and its weight consider the following tokens:

token	frequency	weight
INTERNATIONAL	2183	0.12
HOLDING	1628	0.12
TECHNOLOGY	1207	0.12
AGRIPA	1	1
AAE 1	1	1

The tokens above have been found, for example, in the following strings:

- S1: AAE HOLDING
- S2: AAE TECHNOLOGY INTERNATIONAL
- S3: AGRIPA HOLDING

Their sets of tokens and common tokens are:

---

<sup>5</sup> To illustrate this procedure, consider the following example. “S.F.T. SERVICES SA”, “S.F.T. SERVICES” and “S.F.T. SERVICE” after standardization are transformed into “SFT SERVICE”.

- $t_1 = \{\text{AAE, HOLDING}\}$
- $t_2 = \{\text{AAE, TECHNOLOGY, INTERNATIONAL}\}$
- $t_3 = \{\text{AGRIPA, HOLDING}\}$
- $t_1 \cap t_2 = \{\text{AAE}\}$
- $t_1 \cap t_3 = \{\text{HOLDING}\}$

Without token weighting, strings S1 and S2 have a J distance equal to  $1 - 1/(2+3) = 0.80$ . When the similarity function is adjusted to account for the relevance of each token in the data set the J distance becomes  $1 - 1/(1 + 1.12 + 1 + 1.12 + 0.12) = 0.57$ . In this case weighting reduces the number of operations (and therefore the costs) needed to transform S1 into S2.

#### 4. Results

Our matching experiment focuses on different matching entities: the applicants and applications. Figure 1 reports in the vertical axis the percentage increase in matched names with different similarity levels (J scores) for PATSTAT applicants matched with the 2,197 parent companies and their subsidiaries in our sample. The baseline is the number of matched obtained with a J score of 100% (or 1), corresponding to the maximum level of similarity (perfect match or minimum distance). It is worth to remember that the J score declines with the distance between names and becomes 0 in case of maximum distance. The horizontal axis reports a restricted range of the J score (75% to 100%). The reason why we use a 75 per cent J score as a lower bound is that below this value the quality of the matching, as we show later on, deteriorates very rapidly.

Figure 1 shows that, relative to the baseline (J=100%), the number of applicants matched increases substantially when the level of precision is allowed to decline.

Figure 2 reports the same results for the number of matched applications. The number of matched applications also increases with decreasing levels of the J score. However, the gains relative to the baseline J score are smaller than in the case of matched applicants. The reason for this difference is that many applications are filed by few large patent assignees whose names are often more standardized. Therefore, the potential gains from similarity matching as compared with perfect matching are relatively limited.

It is interesting to note that for both parent applicants and applications the gain in terms of number of matches is greater in the case of US patents than EPO patents – i.e., the relative percentage of matching at the baseline is higher for EPO names. This may be explained by the fact that EPO names and Amadeus names are more similar than USPTO names and Amadeus names in our dataset.

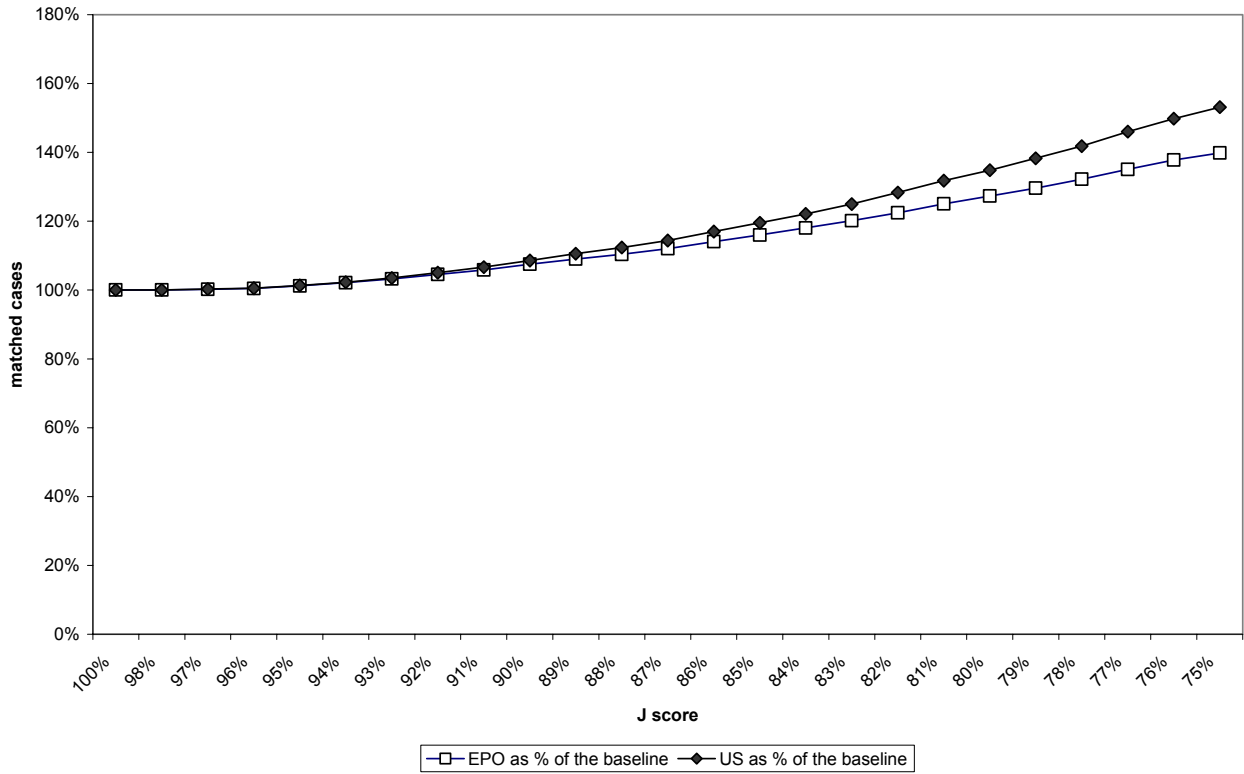


Figure 1 Number of matching links by different level of J score – Applicants

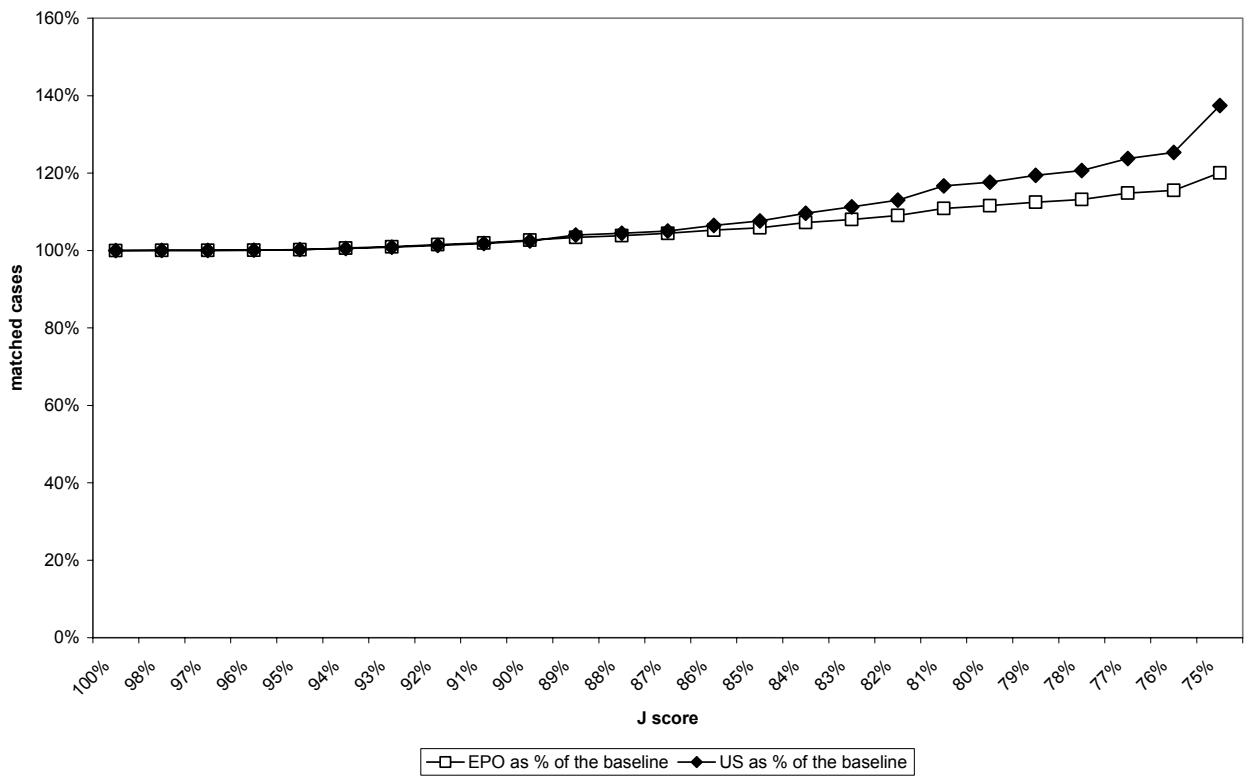


Figure 2 Number of matching links by different level of J score – Applications

Table 2 reports the number of matched patents by sector with a J score larger than 75 per cent.<sup>6</sup> The distribution of matched patents by sector appears to be in line with the distribution of R&D expenditures reported in Table 1, with the exception of pharmaceuticals.<sup>7</sup> The Spearman's rank correlation between R&D and patents by sector is about 0.83 (p-value = .0000).

It is also interesting to see how patents obtained by our matching method correlate with R&D expenditures at the firm level. Figure 3 reports the Pearson's correlation index between the number of patents and R&D expenditures at different levels of the J score. The R&D-patent correlation remains quite stable up to levels of J score of 76% and then declines sharply especially in the case of US patents (and US and EPO patents combined). This result confirms that allowing for lower levels of the J score leads to a substantial loss of precision.<sup>8</sup> Moreover, Figure 3 suggests that the maximum level of patent-R&D correlation is reached at levels of J score between 0.75 and 0.76. Figure 4 digs deeper into the association between patents and R&D at the firm level by showing that the number of patents per R&D expenditures increases at lower levels of the J score. And, in particular, below  $J=0.72$  the patent-R&D ratio bursts up. We should remember that lower levels of the J score imply a higher risk of assigning a patent to the wrong R&D-disclosing firm.<sup>9</sup>

---

<sup>6</sup> We started with 11,903 original applicant names in EPO granted patents and ended up with 1,256 harmonized names.

<sup>7</sup> The small share of patents by pharmaceuticals firms relative to their share of R&D expenditures is in line with the declining R&D productivity of this industry reported by earlier works (e.g., Lanjouw and Shankerman, 2004).

<sup>8</sup> Similarly, the Spearman's ranks correlation (not shown) indicates that for lower levels of J score we have a rapid decrease in the patent-R&D correspondence at the firm level.

<sup>9</sup> Drawing on a subset of the 2,197 firms and harmonized names obtained with the string similarity approach described here, Hall, Thoma and Torrisi (2007) have analyzed the market value of EPO and USPTO patents.

**Table 2 Distribution of matched granted patents by sector with a J score > 0.75**

2.5 digit industry class	EP patents		US patents		EP + US patents	
	n	%	n	%	n	%
01 Food & tobacco	5060	1.9	8071	2.6	13131	2.3
02 Textiles, apparel & footwear	1140	0.4	1672	0.5	2812	0.5
03 Lumber & wood products	140	0.1	75	0.0	215	0.0
04 Furniture	606	0.2	935	0.3	1541	0.3
05 Paper & paper products	2236	0.9	2046	0.7	4282	0.8
06 Printing & publishing	462	0.2	236	0.1	698	0.1
07 Chemical products	34888	13.4	30966	10.1	65854	11.6
08 Petroleum refining & prods	11898	4.6	10947	3.6	22845	4.0
09 Plastics & rubber prods	3281	1.3	3487	1.1	6768	1.2
10 Stone, clay & glass	5186	2.0	5332	1.7	10518	1.9
11 Primary metal products	5972	2.3	8294	2.7	14266	2.5
12 Fabricated metal products	5912	2.3	8560	2.8	14472	2.5
13 Machinery & engines	10647	4.1	12183	4.0	22830	4.0
14 Computers & comp. equip.	1131	0.4	2573	0.8	3704	0.7
15 Electrical machinery	44031	16.9	51938	16.9	95969	16.9
16 Electronic inst. & comm. eq.	19123	7.3	36958	12.0	56081	9.9
17 Transportation equipment	2885	1.1	3551	1.2	6436	1.1
18 Motor vehicles	34932	13.4	50714	16.5	85646	15.1
19 Optical & medical instr	3994	1.5	4078	1.3	8072	1.4
20 Pharmaceuticals	25314	9.7	28516	9.3	53830	9.5
21 Misc. manufacturing	1770	0.7	1483	0.5	3253	0.6
22 Soap & toiletries	9273	3.6	6561	2.1	15834	2.8
24 Computing software	1139	0.4	1852	0.6	2991	0.5
25 Telecommunications	4289	1.6	4462	1.5	8751	1.5
26 Wholesale trade	1034	0.4	778	0.3	1812	0.3
27 Business services	512	0.2	572	0.2	1084	0.2
28 Agriculture	0	0.0	0	0.0	0	0.0
29 Mining	1726	0.7	2624	0.9	4350	0.8
30 Construction	1813	0.7	919	0.3	2732	0.5
31 Transportation services	4631	1.8	3690	1.2	8321	1.5
32 Utilities	6694	2.6	8264	2.7	14958	2.6
33 Trade	277	0.1	245	0.1	522	0.1
34 Fire, Insurance, Real Estate	198	0.1	146	0.0	344	0.1
35 Health services	128	0.0	110	0.0	238	0.0
36 Engineering services	1972	0.8	3179	1.0	5151	0.9
37 Other services	6545	2.5	1271	0.4	7816	1.4
Overall	260839	100.0	307288	100.0	568127	100.0

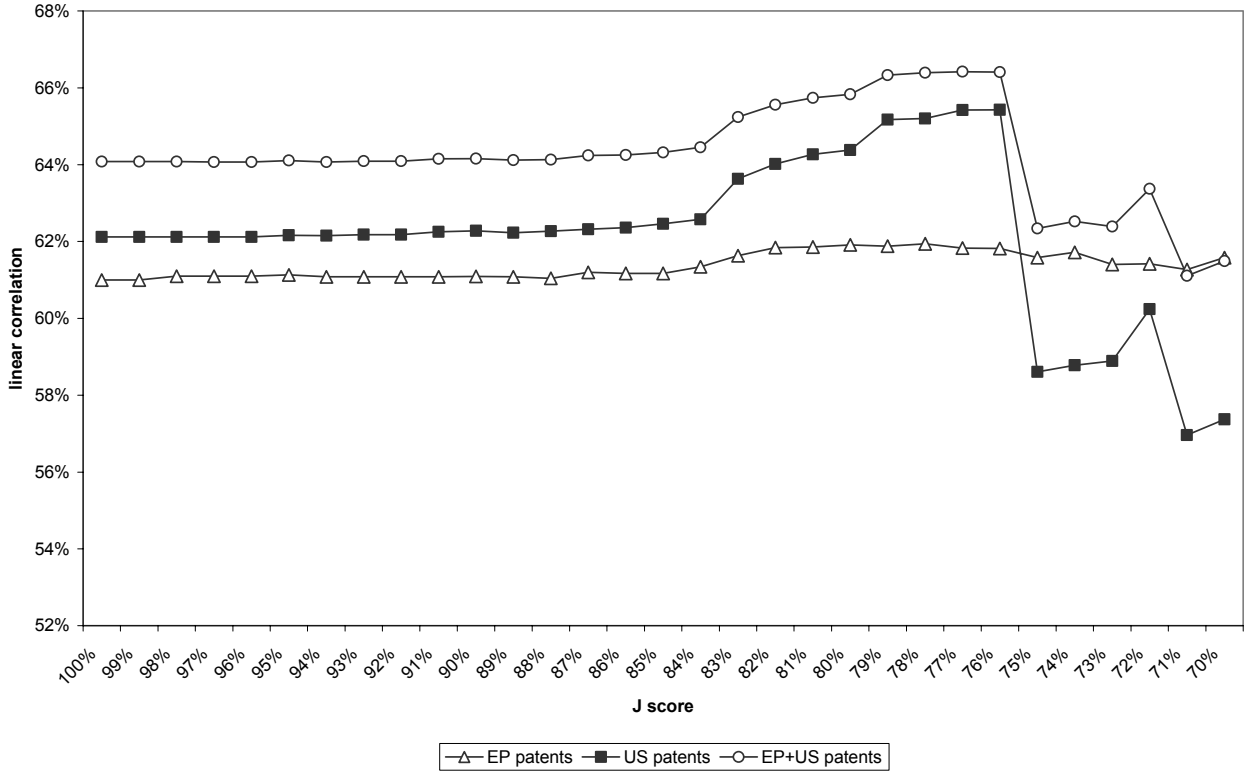


Figure 3 Pearson Correlation Index of R&D and patents by different levels of the J score

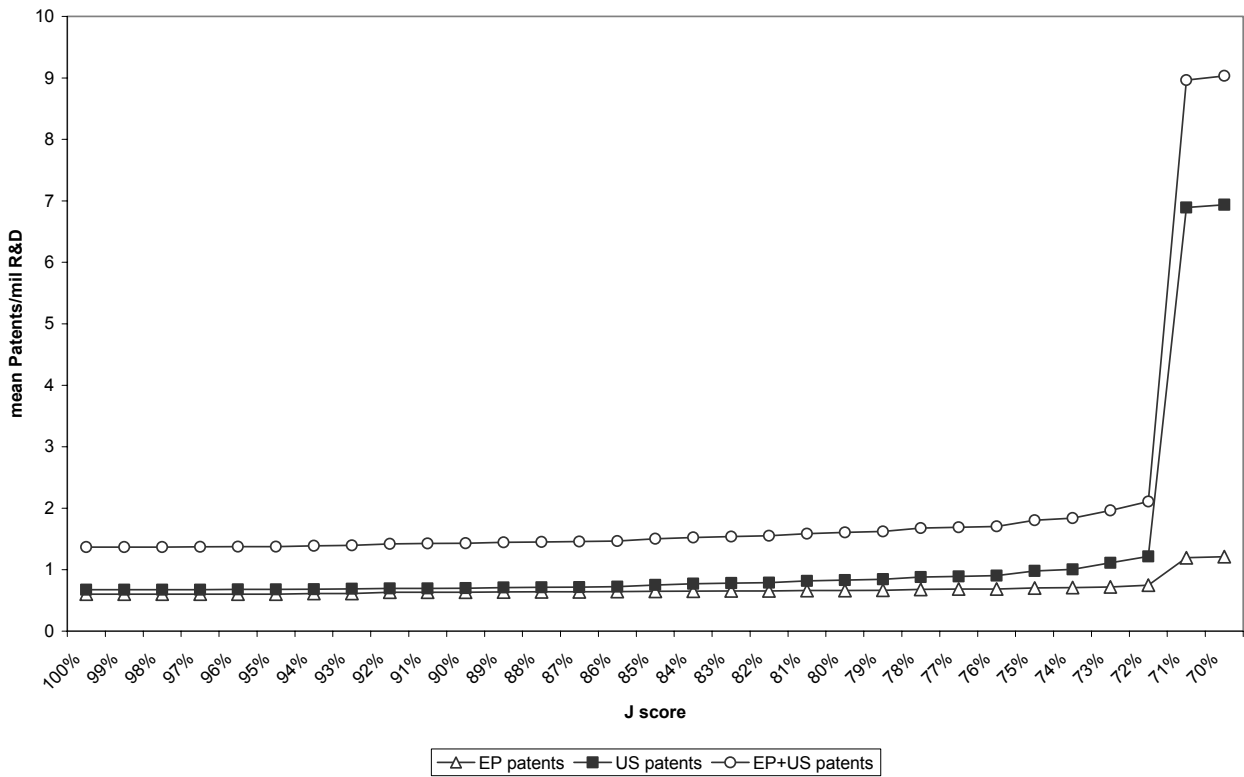
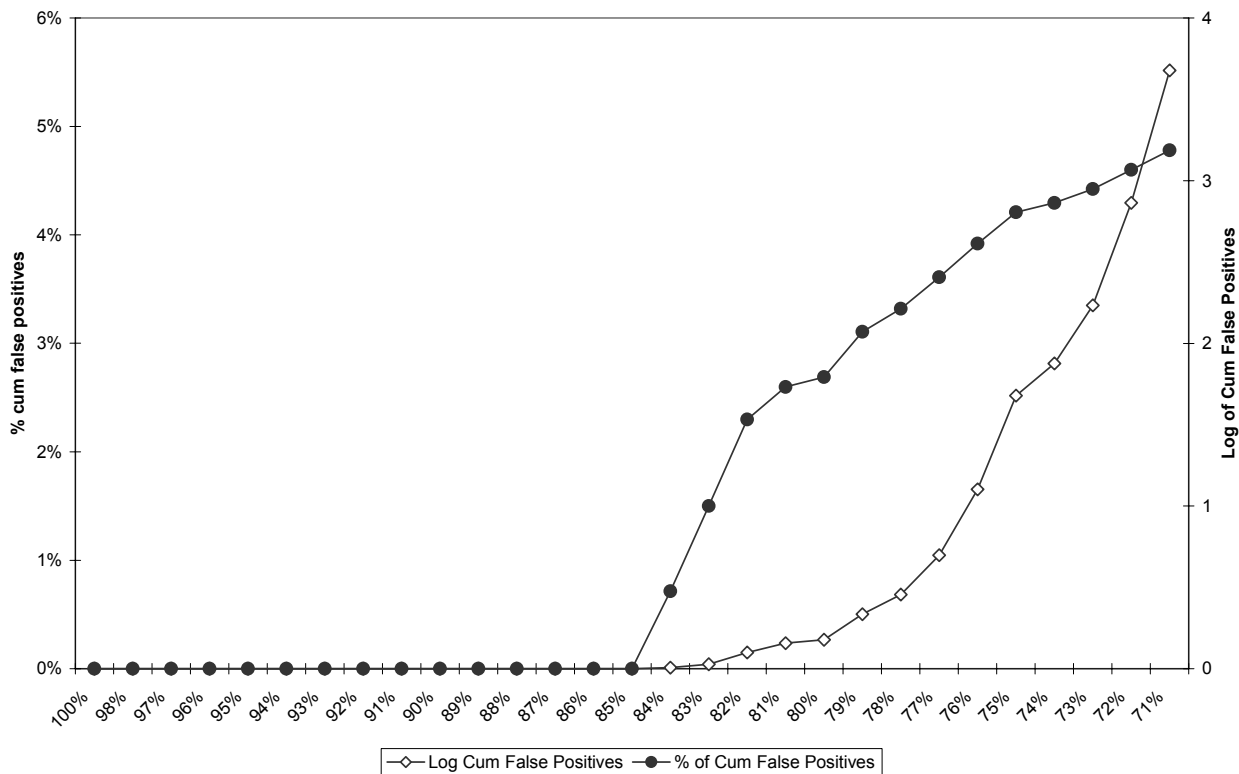


Figure 4 Mean of the ratio patents/R&D by different levels of J score

A different way to find the lowest acceptable level of the J score is to see how the levels of false positives and false negatives vary with the J score.

To estimate the incidence of false positives we focused on EPO patents and checked manually all the occurrences up the level of J score = 70%. As Figure 6 clearly shows, there are small numbers of false positives. The frequency of false positives falls to zero for levels of the J score larger than 85%. In future research we will conduct the same analysis on USPTO patents.

**Figure 5 Cumulative false positives by different levels of J score – EPO patents**



We searched manually for cases of false negatives in the case of EPO patents. To see whether our method fails to match a substantial number of applicants we checked all the European applicants with 15 patents or more. There are 1,326 European applicants falling in this category which have not been matched by our procedure. Only 112 of such cases (8.5%) can be considered as false negatives. A large share of false negatives is due to differences in the applicant address between PATSTAT and Amadeus. Other false negatives are due to spelling errors and missing tokens in company names.

### 5. Robustness checks

In this section we compare our results with those obtained by other standardization methods. In particular, we consider as a benchmark the Thomson Scientific’s Derwent World Patent Index (2002).

The Derwent Index covers about 21,000 assignees. Each assignee is given a four-letter code, which is normally based on the name of the applicants. Prior to 1992 a maximum of four applicants per patent document were assigned a code. From 1963 to 1969 all applicants, including individuals, were assigned four-letter codes. After 1970 unique codes have been assigned to companies who make a significant number of patent applications. These companies and their four letter codes are named ‘standard’ while other companies are treated as ‘non-standard’. The subsidiaries of large groups are normally assigned the same standard code, even when their names differ from that of the

parent company. For example the code PENN is used for the following list of firms belonging to the same legal entity:

- Pennsalt Chem Corp
- Pennsylvania Salt Mfg Co
- Pennwalt Corp
- Pennwalt France SA
- Pennwalt Holland BV
- Pennwalt Ltd

In cases of conglomerates, like the Japanese Mitsubishi, Toshiba and Hitachi, individual subsidiaries may be given their own codes.

To maintain a given level of consistency over time, in case of change of company names Derwent retains the standard code. For example, Bayer AG, formerly Farbenfabriken Bayer AG, is still coded FARB. When two organizations, with standard patent assignee codes, merge Derwent normally maintain the standard patent assignee code for each organization as long as patents filed under the names of the independent organizations continue to appear.

For instance, the SANO and CIBA codes have continued to be applied to Novartis (NOVS) after the merger of Sandoz (SANO) and Ciba (CIBA) for all patents filed under the names of Sandoz and Ciba. However, in case of M&As, demergers and takeovers that involve two large companies Derwent does not follow a standard procedure. If a new code was generated for Novartis (merger), in other cases one code was maintained and the other was dropped – e.g., Smithkline Beecham, Bristol-Myers Squibb and Glaxo Wellcome.

Finally, applicants codes are not generally changed retrospectively.

Although the Derwent standardization procedure was developed for US patent assignees, it can also be applied to other datasets like Amadeus and EPO.

We standardized applicant names in PATSTAT and Amadeus according to the Derwent index and used the results of this procedure as a benchmark for our standardization method.<sup>10</sup>

---

<sup>10</sup> Rachel Griffith, Gareth Macartney and colleagues at the IFS have developed a software implementation of Derwent procedure. They have also implemented some standard cleaning and punctuation removal to the ASCII standard code. We thank these colleagues for kindly providing us with the STATA code.

**Table 3 Distribution of matched granted patents by sector with the Derwent method as share our matching**

2.5 digit industry class	EP patents %	US patents %	EP + US patents %
01 Food & tobacco	40.7	69.0	58.1
02 Textiles, apparel & footwear	826.9	91.9	389.9
03 Lumber & wood products	99.3	76.0	91.2
04 Furniture	443.2	72.3	218.2
05 Paper & paper products	33.9	53.9	43.4
06 Printing & publishing	2.4	67.8	24.5
07 Chemical products	82.8	74.8	79.0
08 Petroleum refining & prods	57.6	81.0	68.8
09 Plastics & rubber prods	62.6	63.0	62.8
10 Stone, clay & glass	46.5	61.4	54.0
11 Primary metal products	56.0	40.7	47.1
12 Fabricated metal products	50.3	74.5	64.6
13 Machinery & engines	81.1	81.5	81.3
14 Computers & comp. equip.	47.5	56.4	53.7
15 Electrical machinery	49.8	92.7	73.0
16 Electronic inst. & comm. eq.	90.0	47.6	62.0
17 Transportation equipment	73.5	53.5	62.5
18 Motor vehicles	69.8	88.6	80.9
19 Optical & medical instr	63.3	68.9	66.1
20 Pharmaceuticals	66.5	86.3	77.0
21 Misc. manufacturing	63.9	95.0	78.1
22 Soap & toiletries	66.7	62.2	64.9
24 Computing software	90.7	65.0	74.8
25 Telecommunications	87.2	80.5	83.8
26 Wholesale trade	39.3	75.7	54.9
27 Business services	55.7	67.3	61.8
28 Agriculture	na	na	na
29 Mining	118.8	35.2	68.4
30 Construction	61.6	54.4	59.2
31 Transportation services	63.2	89.1	74.7
32 Utilities	75.8	68.2	71.6
33 Trade	135.4	83.7	111.1
34 Fire, Insurance, Real Estate	66.7	63.0	65.1
35 Health services	71.1	77.3	73.9
36 Engineering services	59.5	29.3	40.9
37 Other services	3.1	25.9	6.8
Overall	69.7	75.2	72.6

A first level of comparison concerns the total number of matched patents. Table 3 shows the sectoral distribution of patents matched by the Derwent method. We should recall that the Derwent method is used to carry out perfect matches between company names in PATSTAT and Amadeus by relying on the standard four digit codes assigned by Derwent. This is different from the case of J score=100%, which is calculated by weighting all tokens in each string. The use of approximate matching algorithms can yield a gain of around 40 % over the perfect name matching, with both US patents and EPO patents. However the matching gain varies significantly across sectors. In traditional sectors, such as textiles, apparel & footwear, furniture, mining and trade, the Derwent method outperforms the approximate matching. This suggests that the perfect matching method is better at tracing the evolution of company names in traditional sectors. But this issue should be

examined more carefully because the standardized names reported by the Derwent Index are not unique for a given company and this may give rise to a substantial number of false positives. Moreover, in sectors characterized by higher turbulence (large numbers of entries and exits, and M&As), such as computers and telecommunications, approximate matching has a better performance than perfect matching.

A further comparison between the two methods can be done on the ground of accuracy. First, we found that around 89.9% of patents matched by the Derwent method are also matched by our procedure. Second, about 94.2% of applicants matched by the Derwent method are also matched by our method. Third, 82.3% of patents-applicants matched by the Derwent method are also matched by the approximate matching procedure. However, using the Derwent method leads to 314 cases where the number of matched legal entities (from Amadeus) is larger than the number of applicants (from PATSTAT). By contrast, the approximate matching yields only 29 of such cases. These numbers may point out a higher accuracy of the ASM method as compared with the Derwent file. A more accurate analysis of false positives generated by the Derwent method will be done in future research.

## 5. Further sources of name standardization: exploiting the priority links between USPTO and EPO patent databases

In this section we analyze an additional standardization method for applicant names using priority links between USPTO and EPO patent databases. The objective of this analysis is to see whether the priority links between US and EPO patents can improve the accuracy of standardization and therefore have substantial positive effect on the quality of the similarity matching procedure use in our exercise.

The reason why we conduct this exploration is that the USPTO provides a list of standardized assignee names. These standardized names can be downloaded from the NBER patent database ([www.nber.org/patents/](http://www.nber.org/patents/)). The file collects information on all the applicants that have been granted at least one patent by the USPTO over the period 1963-2002.

Our name standardization process exploits the priority links between all EPO patent applications and all USPTO granted patents by following five steps reported in Figure 8. We include all EPO patent applications in the standardization process with the objectives of standardizing as much documents as possible.

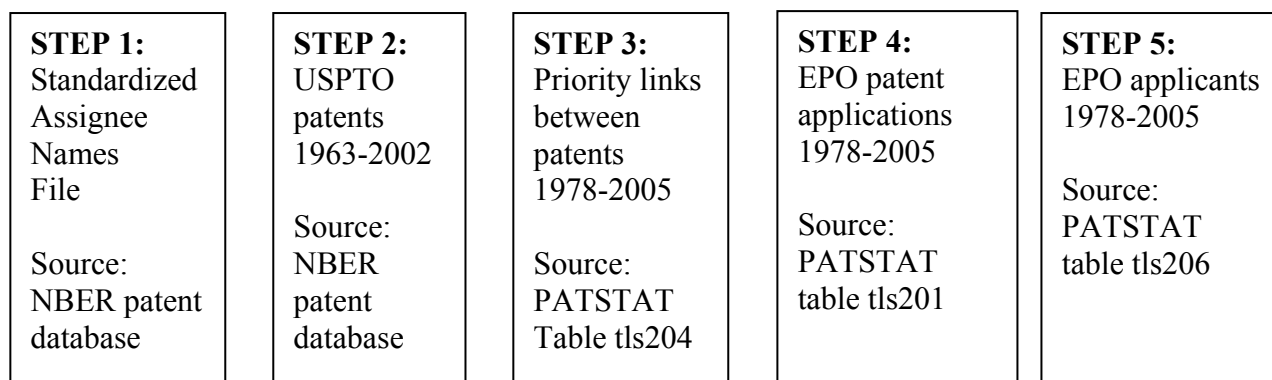


Figure 6 Standardization Process, Data and Sources

### STEP 1.

**coname** is the label for the file of Standardized Assignee Names in the USPTO system. This dataset includes 203,331 distinct assignees names.

**Table 4 Size Distribution of Assignees in USPTO 1969-2002 (with Standardized Name)**

patents	assignee	Percent
1	109387	53,8
2-5	61954	30,47
6-15	19241	9,46
16-99	10319	5,07
100-499	1846	0,91
500-999	294	0,14
1000+	290	0,14
<b>Total</b>	<b>203331</b>	<b>100</b>

**STEP 2.**

pat63\_02f is the file with the patent number and assignee number from the NBER patent database. It includes information on the front-page of the patent for around 3,416,957 granted patents. For 2,163 patents the name of the assignee was missing.

**STEP 3**

Using PATSTAT database we build up the priority links (patent family) of all EPO patents by working on Table t1s204\_appln\_prior in PATSTAT. This table reports for each patent the application\_id and the priority\_application\_id. Drawing on the priority links, the patent family can be defined as follows. First, we have isolated all priorities reported by EPO patents. This yields around 2,168,406 combinations of EPO patents and their priorities (remember that we have a many-to-many correspondence between patents and priorities). Then, we need to find the non-EPO patents which report the same priority as the EPO patents above which will provide additional correspondences. These two queries should provide a first subset of the patent family. We obtained a set composed of 15,237,134 correspondences.

Secondly, we need to find all applications (EPO and non-EPO) which report an EPO application as priority. This step should lead to the second subset of the family size (about 555,110 correspondences). Obviously, we accounted for double counting of correspondences contained in both subsets.

Once the patent family for EPO patents has been built up we need to isolate the priority links that involve at least one USPTO patents retrieved in Step 2. We could isolate 1,158,634 priority links between USPTO and EPO, involving 1,008,243 US granted patents and 942,273 EPO applications.

**STEP 4**

In this step we used a database compiled by from PATSTAT by linking each application with their priorities in US patents.

**STEP 5**

The identification of the proper link to the EPO applicants has been executed taking into account the number of priority links and the number of applicants per EPO patents. Table 5 illustrates four situations that characterize the dataset.

**Table 5 Priority links to USPTO patents by number of applicants in EPO patents**

applicants\Priority links	One priority link per patent	More priority links per patent	Overall
<b>One applicant per patent</b>	776,817	105,729	882,546
<b>More applicants per patent</b>	51,911	7,795	59,706
<b>Overall</b>	<b>828,728</b>	<b>113,524</b>	<b>942,252</b>

It is worth to recall that the standardization process is limited by two constraints. First, when for a given EPO patent there are two or more applicants whereas there is one priority link (one assignee name) it is not possible to establish automatically a proper US-EPO name correspondence. This is because assignee names in STEP 1 and applicant names in STEP 5 may be different and often the original order of occurrence has not been recorded. Hence, the identification of the proper links can be done automatically only when we consider single applicant patents in EPO and USPTO and one-to-one priority links between EPO and USPTO patents (77,817 cases in Table 5). In 51,911 cases this is not the case and then we cannot identify automatically which Standardized Assignee Names in Step 1 corresponds to which EPO applicant in Step 5. Second, when in STEP1 we find many-to-many priority links (more assignee names per patent) we need to apply some additional testing to avoid false matches (113,524 cases). At this explorative stage we conducted a manual testing of the standardization process.

We started by isolating all granted patent applications excluding withdrawals (the average quality and consistency of data contained in these documents is limited). Then we focused only on the documents having one applicant and one priority link to a single assignee patent in USPTO, and restricting to business applicants only.<sup>11</sup> Starting from 776,817 links then we ended up with a subset of 373,813 patent documents or 24.37% of the total EPO applications originated by business applicants. This subset involves around 35,466 applicants that can be associated uniquely and automatically to an USPTO Standardized Assignee Name. In case of inconsistency between two documents having two distinct USPTO Standardized Assignee Names while in the EPO equivalents they have the same applicant, we have standardized to the USPTO Assignee Name having more patents in USPTO.<sup>12</sup> These 35,466 standardized applicant names account for about 81.80% of all EPO applications filed by business organizations.

For the remaining applications with priority links we checked manually the association between USPTO names and EPO names and obtained an additional 20% (7,083) US standardized names, accounting for 3.28% of all EPO applications filed by business organizations.

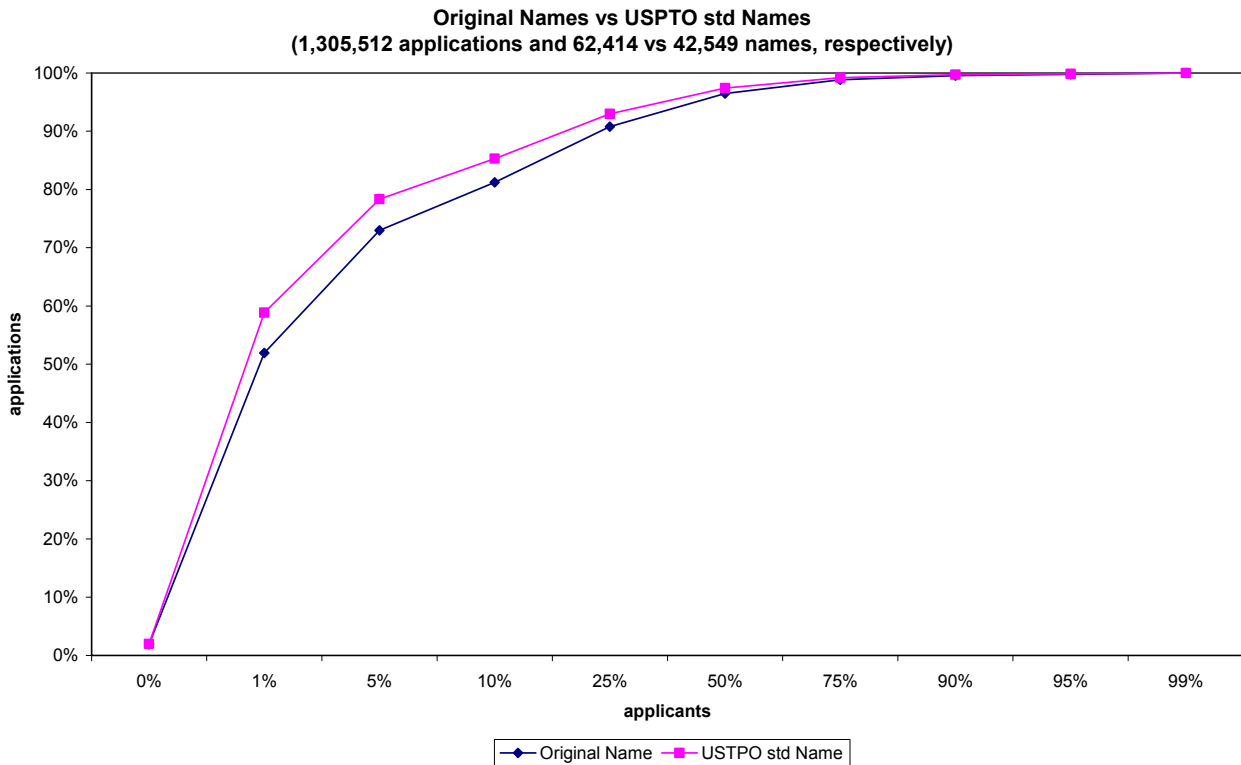
The final list of EPO applicants with a US standardized name includes 42,549 units corresponding to 85.06% of all applications filed by business organizations to the EPO.

Figure 7 shows the gains in terms of standardization with respect to the original source names. The horizontal axis shows the percentiles of assignees while in the horizontal axis is reported the cumulative share of patents. Starting from 62,414 original names of EPO patent applicants we reached about 42,549 USPTO standard names – a reduction of about 43 per cent.

---

<sup>11</sup> The data on the number of applicants have been provided by Colin Webb from OECD patent repository.

<sup>12</sup> In such cases it would be better, from a methodological point of view, to rely on string similarity functions to establish the link between USPTO and EPO names. However, the marginal benefits of this approach would be limited because of the high computation costs required and the small number of cases involved.



**Figure 7 Gains from using USPTO standardized names**

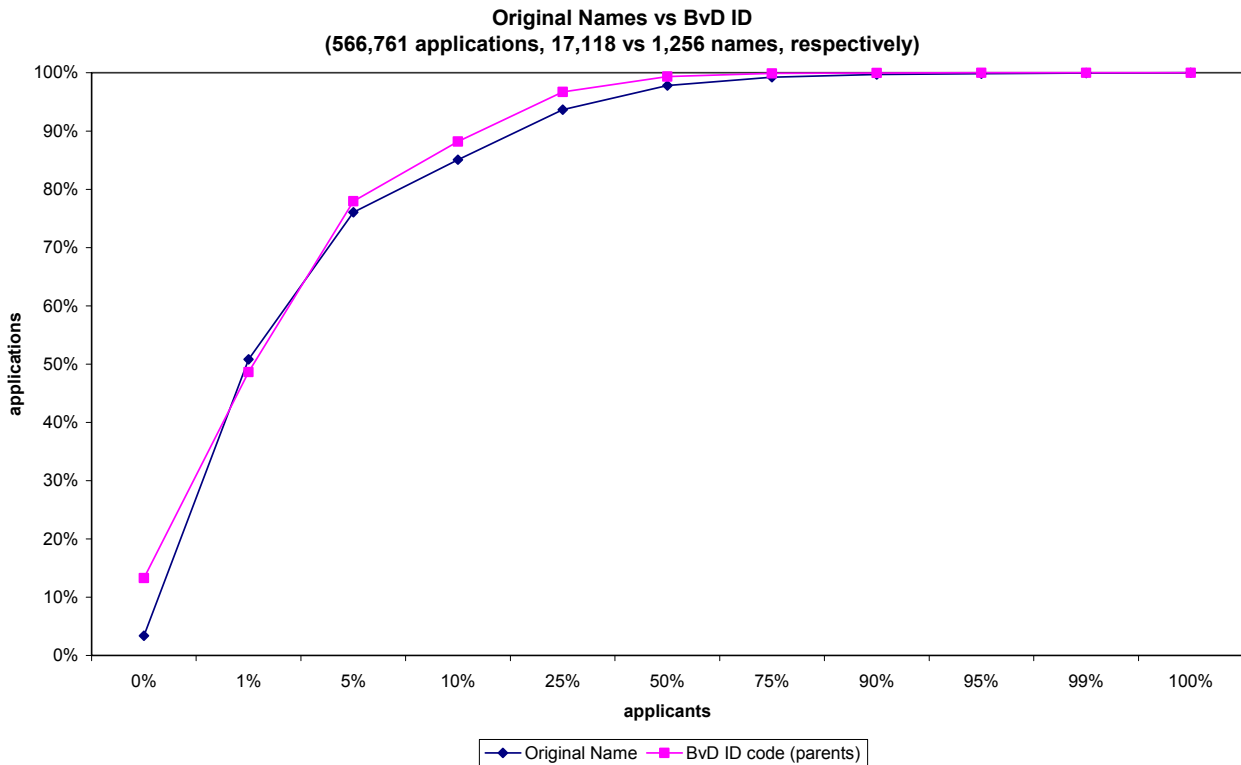
The benefits of standardization with US-EPO priority links illustrated above leads us to replicate the string similarity matching exercise by using USPTO CONAMES for our 2,197 sample firms. Figure 8 reports the gains from string similarity matching without using USPTO CONAMES. The EPO dataset contains 566,761 applications and 17,118 original applicants.<sup>13</sup>

The matching based on J score leads us to find 1,256 harmonized names. Figure 9 reports the gains obtained by our similarity matching method when USPTO conames are added in the name standardization stage. Thanks to the priority links with the USPTO standardized assignee names, we found 641,423 EPO patent applications for our sample firms (instead of the 566,761 patent applications founded without exploiting the priority links with the USPTO names). These applications correspond to 20,074 non standardized EPO applicant names which are transformed into 6,847 names when the US standardized names are used. The number of standardized names falls to 1,256 names when the information on corporate structure provided by Amadeus (BVD numbers) is also taken into account.

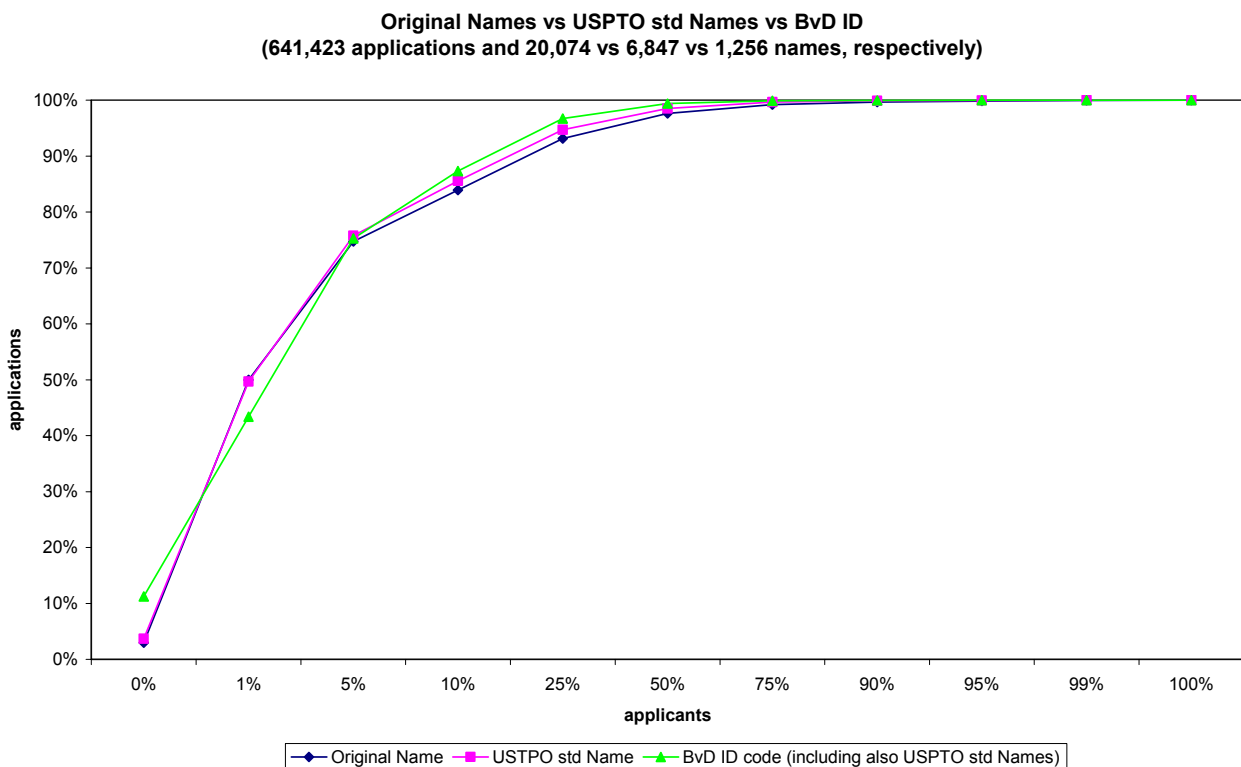
The gain from using conames in our matching method in terms of a larger number of patent applications matched is then substantial. This gain is higher for the medium-sized assignees as demonstrated by the declining difference between the two distributions between the 10<sup>th</sup> and 25<sup>th</sup> percentiles.

This test shows that approximate matching algorithms to business directories can substantially benefit from the use of standardized USPTO assignees names. The use of US standardized names can also improve the precision of standardization by taking into account changes in firm names and ownership links that have not been recorded in business directories like Amadeus.

<sup>13</sup> Considering EPO granted patents only we have 260,839 applications and 11,903 original applicants.



**Figure 8 Gains from string similarity matching without USPTO conames**



**Figure 9 Gains from using USPTO standardized names and string similarity matching**

## 6. Conclusions

This paper illustrates the results of a test of company name standardization and matching. Our analysis is based on two data sources: the PATSTAT patent database (USPTO and EPO patents)

and the Amadeus accounting and financial dataset. Our experiment focuses on 2,197 European publicly listed firms and their subsidiaries. Earlier studies have mostly relied on manual, ad-hoc methods. More recently some scholars have started experimenting with automatic matching techniques. Our results contribute to the literature by comparing two different approaches – the character-to-character match of harmonized company names (perfect matching) and the approximate matching based on the J similarity string index. Our results show that approximate matching yields substantial gains over perfect matching, in terms of frequency of positive matches, with a limited loss of precision – i.e., low rates of false positives and false negatives. Moreover, our study shows that for levels of J similarity score above 75 per cent the distribution of matched patents by sectors is similar to the sectoral distribution of the R&D expenditures. Moreover, patent-R&D ratios in specific sectors appear to be in line with those reported in studies on R&D productivity. Furthermore, the correlation between R&D expenditures and patents at the firm level remains high and significant for J scores between 75 and 100 per cent. Below the 75 per cent level, the number of matched patents per R&D expenditure increases very rapidly mostly because of an increasing imprecision in names matching. Our analysis also shows that the number of false positives is linearly decreasing in the level of the Jaccard similarity score. In future research we will use confidence intervals derived from the distribution of false positives to weight our matching links. Finally, we have used the USPTO standardized assignees names (conames) to find the priority links between EPO patents and USPTO patents. Our analysis shows that accounting for these priority links improves the performance of names matching based on the J similarity string index.

## References

- Arundel, A. (2003), *Patents in the Knowledge-Based Economy*, Report of the KNOW Survey, MERIT, University of Maastricht.
- Arora, A., Fosfuri, A. and Gambardella, A., (2003), "The Division of Inventive Labor: Functioning and Policy Implications", Paper presented at the CREST conference in honour of Zvi Griliches, Paris August 25-27, 2003
- Cohen, W. M., R. R. Nelson, et al. (2000). Protecting Their Intellectual Assets: Appropriability Conditions and Why U.S. Manufacturing Firms Patent (or Not). *NBER Working Paper* No. 7552. Washington, DC, NBER.
- Derwent (2000), *WORLD PATENTS INDEX - Derwent Patentee Codes*, Revised Edition 8 ISBN: 0 901157 38 4, Thomson Publishers. Leuven Manual.
- Fosfuri, A., and Giarratana, M.S. (2007). Product Strategies and Survival in Schumpeterian Environments: Evidence from the US Security Software Industry. *Organization Studies* 28 (6): 909-929.
- Gambardella, A., D. Harhoff, and B. Verspagen (2005), "The Value of Patents." Universita Bocconi, Ludwig-Maximilians Universitaet, and Eindhoven University, Working Paper: [http://www.creiweb.org/activities/sc\\_conferences/23/papers/gambardella.pdf](http://www.creiweb.org/activities/sc_conferences/23/papers/gambardella.pdf)
- Giarratana, M. and Torrisi, S. (2004.), "Entry and Survival in Foreign Markets: Technology, Brand Building and International Linkages", *Social Science Research Network - Electronic Paper Collection*, SSRN\_ID577401\_code386435.pdf (<http://papers.ssrn.com>).
- Giuri, P., Mariani, M. et al. (2005) "Everything you Always Wanted to Know about Inventors (but Never Asked): Evidence from the PatVal-EU Survey". LEM Papers Series 2005/20, Sant'Anna School of Advanced Studies, Pisa, Italy.
- Griliches, Z. (1981), "Market Value, R&D and Patents." *Economic Letters* 7: 183-87.
- Griliches, Z. (1990), Patent Statistics as Economic Indicators: A Survey, *Journal of Economic Literature*, XXVIII (Dec.): 1661-1707.

- Griliches, Z., Hall, H. B. and Pakes, A. (1991), "R&D, Patents. And Market Value Revisited: Is There a Second (Technological Opportunity) Factor?." *Economics of Innovation and New Technology* 1: 183-202.
- Hall, B. H., A. Jaffe, and M. Trajtenberg (2001), "The NBER Patent Citations Data File: Lessons, Insights, and Methodological Tools." In A. Jaffe and M. Trajtenberg (eds.), *Patents, Citations and Innovations*, Cambridge, MA: The MIT Press. Also Cambridge, Mass.: National Bureau of Economic Research Working Paper 8498 (October).
- Hall B. H., A. Jaffe, and M. Trajtenberg (2005), "Market Value and Patent Citations," *Rand Journal of Economics* 36: 16-38.
- Hall H. B., Thoma G. and Torrisi S. (2007) The market value of patents and R&D: Evidence from European firms, Working paper 13426, National Bureau of Economic Research, Cambridge, Mass. (<http://www.nber.org/papers/w13426>).
- Harhoff, D., F. Narin, and K. Vopel (1999), "Citation Frequency and the Value of Patented Inventions." *Review of Economics and Statistics* 81(3): 511-15.
- Jaccard, P. (1901) Bulletin del la Société Vaudoisedes Sciences Naturelles 37, 241-272.
- Lanjouw, J. O., and M. Schankerman (2004), "Patent Quality and Research Productivity: Measuring Innovation with Multiple Indicators." *Economic Journal* 114: 441-465.
- Levenshtein, V.I. (1966) Binary codes capable of correcting deletions, insertions, and reversal, *Soviet Physics Doklady*, 10(8) S. 707-710.
- Levin, R. C., A. K. Klevorick, et al. (1987). "Appropriating the Returns from Industrial Research and Development." *Brooking Papers on Economic Activity* 3: 783-831.
- Magerman, T. Van Looy B., and Song X. (2006) Data production methods for harmonized patent statistics: Patentee name standardization. Technical report, K.U. Leuven FETEW MSI.
- Moser, P. (2005) How Do Patent Laws Influence Innovation? Evidence from Ninetheenth-Century World Fairs, *The American Economic Review*, vol. 95 (4), September, pp. 1215-1236
- Navarro, G. (2001). "A guided tour to approximate string matching". *ACM Computing Surveys* 33 (1): 31--88.
- Griliches, Z. (1990) Patent Statistics as Economic Indicators: A Survey, *Journal of Economic Literature*, Vol. XXVIII, December 1990, 1661-1707.
- Patel, P. and K. Pavitt (1991). "Large firms in the production of the world's technology: an important case of 'non-globalisation'." *Journal of International Business Studies* 22 (1), 1-21.
- Pavitt, K. (1985) 'Patent Statistics as an Indicator of Innovative Activities: Possibilities and Problems', *Scientometrics*, 7 (1-2): 77-99.
- Pavitt, K. (1988) "Uses and abuses of patent statistics," in van Raan, A. (ed.) *Handbook of Quantitative Studies of Science Policy*, Amsterdam: North Holland.
- Pavitt, K., Robson, M. and Townsend, J. (1987), 'The Size Distribution of Innovating Firms in the UK: 1945–1983', *Journal of Industrial Economics*, March, 35, 291–316.
- Powell, W. W., D. R. White, et al. (2005). "Network Dynamics and Field Evolution: The Growth of Interorganizational Collaboration in the Life Sciences." *American Journal of Sociology* 110(4): 1132-1205.
- Salton, G. and Buckley, C. (1988) Term-weighting approaches in automatic text retrieval. *Information Processing & Management* 24(5): 513–523.
- Schmookler J. (1966) *Invention and Economic Growth*, Harvard University Press, Cambridge, MA
- Smith, T. F. and Waterman, M.S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology* 147: 195-197.