Institut de Recerca en Economia Aplicada Regional i Pública
Research Institute of Applied Economics

Document de Treball 2011/05 39 pàg
Working Paper *2011/05 39 pag.*

"Singling out individual inventors from patent data"

Ernest Miguélez and Ismael Gómez-Miguélez

UB IREA

de Recerca en Economia Aplicada | *Research Institute of Applied Economics*

Institut de Recerca en Economia Aplicada Regional i Pública
Research Institute of Applied Economics
Universitat de Barcelona
Av. Diagonal, 690 • 08034 Barcelona

WEBSITE: www.ub.edu/irea/ • CONTACT: irea@ub.edu

The Research Institute of Applied Economics (IREA) in Barcelona was founded in 2005, as a research institute in applied economics. Three consolidated research groups make up the institute: AQR, RISK and GiM, and a large number of members are involved in the Institute. IREA focuses on four priority lines of investigation: (i) the quantitative study of regional and urban economic activity and analysis of regional and local economic policies, (ii) study of public economic activity in markets, particularly in the fields of empirical evaluation of privatization, the regulation and competition in the markets of public services using state of industrial economy, (iii) risk analysis in finance and insurance, and (iv) the development of micro and macro econometrics applied for the analysis of economic activity, particularly for quantitative evaluation of public policies.

Abstract

An increasing number of studies in recent years have sought to identify individual inventors from patent data. A variety of heuristics have been proposed for using the names and other information disclosed in patent documents to establish "who is who" in patents. This paper contributes to this literature by describing a methodology for identifying inventors using patents applied to the European Patent Office (EPO hereafter). As in much of this literature, we basically follow a three-step procedure: (1) the parsing stage, aimed at reducing the noise in the inventor's name and other fields of the patent; (2) the matching stage, where name matching algorithms are used to group similar names; and (3) the filtering stage, where additional information and various scoring schemes are used to filter out these similarly-named inventors. The paper presents the results obtained by using the algorithms with the set of European inventors applying to the EPO over a long period of time.

JEL classification:. C8, J61, O31, O33, R0.
Keywords: "Names game", patent data, unique inventors, name matching algorithms

Ernest Miguélez is Predoctoral Scholar (FPU), Ministry of Education, at University of Barcelona. Department of Econometrics , Statistics and Spanish Economy. E-mail: emiguelez@ub.edu

Ismael Gómez-Miguélez is Predoctoral Scholar (FPI), Ministry of Science and Innovation, at Technical University of Catalonia. Signal Theory and Communications Department. E-mail: ismael.gomez@tsc.upc.edu

Acknowledgements:

Part of this work was carried out while Ernest Miguélez was visiting the Kiel Institute for the World Economy (Kiel, Germany) and the 'Knowledge, Internationalization and Technology Studies' (KITeS) Research Group at Bocconi University (Milan, Italy). The use of their facilities is gratefully acknowledged. Ernest Miguélez received financial support from the Ministerio de Ciencia e Innovación, ECO2008-05314 and AP2007-00792, and from the European Science Foundation, for the activity entitled 'Academic Patenting in Europe'. The usual disclaimer applies.

1. Introduction

Patent data offer a wide range of information for research in innovation economics, regional economics, and economic geography, among other fields in the social sciences. Patent documents contain information about the inventor's name, the owner's[1] name, the year and exact date of application[2], the exact addresses of both the inventor(s) and the applicants, and the technological class to which the patent belongs. Further, merging these datasets with patent citations, non-patent citation literature, and firm data provides even more information and sheds light on the ways in which knowledge is produced, exploited, and spread.

Patent data should be treated with caution, since not all inventions are patented, not all inventions have the same economic impact, and not all patented inventions are commercially exploitable innovations (Griliches, 1991). Nonetheless, patent data have proved their usefulness for proxying inventive activity because they present the minimal standards of novelty, originality and potential profits (Bottazzi and Peri, 2003).

What has been less studied so far is the inventor herself: her personal characteristics, her linkages with other inventors or firms, and her professional and geographical mobility, and the implications of her presence in a given location for regional and national innovativeness, capability, and growth.

The reason why this literature is less fertile is basically that patent data do not provide a consistent list of unique personal identifiers. Unique IDs for each inventor and for anyone else are missing. The information which is closest to being a sort of inventor's ID is her own name (name, middle name, surname, and so on), and for this reason attempts to identify single inventors have mainly used it as a point of departure. However, this procedure is problematic for two main reasons. First, names and surnames contained in the patent document may well be spelled differently in each patent. Second, it is possible that two patents, with exactly the same name (say, John Smith) do not belong to the same inventor.

A large body of literature has sprung up in recent years to deal with these and related problems (Fleming et al., 2007; Carayol and Cassi, 2009; Giuri et al., 2007; Hoisl, 2006; Kim et al., 2006; Lai et al., 2009, Lissoni et al., 2006; Raffo and Lhuillery, 2009; Trajtenberg et al., 2006; Thoma

[1] The owner of a patent is the firm, institution, or individual who appears as the owner in the patent document – under the head "applicant". In the present paper we use the terms *owner*, *applicant*, or *assignee i*ndistinctively.
[2] The priority year is the first year a patent was applied for worldwide.

and Torrisi, 2007).[3] These authors have tried to contribute to the identification of individual inventors by using their names, certain patent characteristics, and different ad-hoc heuristics, in what they called "the Names Game" (Trajtenberg et al, 2006; Raffo and Lhuillery, 2009). So far, however, no one methodology has shown its superiority over the others: indeed, most approaches present new advantages, but a number of shortcomings as well. Our suggestions in the present inquiry are inspired by this earlier literature, and try to contribute to enrich it at the same time. Our aim here will be to exploit what, in our opinion, constitute the main advantages of these studies and at the same time to avoid their main drawbacks. The methodology developed will be applied first to a small sample of inventors which we will use as benchmark to test the goodness-of-fit of the approach, and then to a large dataset of European patents applied for by European inventors over a long period of time.

We should mention that some of the researchers mentioned have recently joined the "Academic Patenting in Europe (APE-INV)" project led by KITES-Bocconi University. This project aims to compile a set of best practices for identifying inventors from patent data. A summary of this project can be found in Lissoni et al. (2010)[4], which also provides an updated survey of related studies.

In the next section, we present a detailed explanation of the problems faced and the solutions adopted. Broadly speaking, the aforementioned literature divides the procedure for identifying inventors into three main stages (see Raffo and Lhuillery, 2009). The first stage deals with data cleaning, homogenization and standardization. The second stage matches the name of the inventors in order to form groups of patents potentially belonging to the same inventor. Finally, within each group of patents, a variety of heuristics and algorithms have been used to perform pair-wise comparisons and to establish whether pairs of patents belong to the same inventor or not.

The outline of the paper is as follows. In section 2 we explain in detail the three-step methodology. Section 3 presents some results of the algorithm applied to a subsample of European patents, manually checked by Carayol and Cassi (2009). Section 4 shows the results of applying the methodology to the full list of patent applications presented to the EPO by inventors residing in Europe (EU-27 plus Iceland, Liechtenstein, Norway, and Switzerland) and stored in the REGPAT database (OECD, January 2010 edition), while section 5 concludes and suggests directions for future research.

[3] A brief summary of the different methodologies applied in these studies and the scope of their empirical application is included in the appendix.
[4] See the APE-INV project website: http://www.esf-ape-inv.eu/.

2. The "*Names Game*" using patent data

Patent data contain a huge amount of information that are very useful for a variety of analyses. However, they do not provide a consistent list of unique inventors' personal identifiers. In this situation, it is necessary to use the inventor's name and surname reported in the patent itself. Unfortunately, this strategy faces two main problems. The first occurs when the name (or surname) of the same inventor is spelled differently on different occasions (Ericsson *versus* Eriksson; Webber versus Weber; Smith *versus* Schmyt; and so on). The second concern is known in the literature as "the John Smith problem": i.e. when two inventors with exactly the same name are not actually the same inventor. To cope with this difficulty, the literature suggests performing a list of algorithms aimed to identify single inventors by using their names and surnames and other useful information disclosed in the patent document. Following Raffo and Lhuillery (2009), and using their terminology, we divide the methodology to obtain the final data into three steps: parsing, matching, and filtering stages.

The parsing stage

The first step is to clean up the fields of the database containing the name and surname of the inventor, and the field with their addresses. We also want to homogenize and standardize the structure of each field and its content as far as possible, in order to allow comparisons between records.

For the case of the "inventor's name" field, basically we proceed in two ways. First, following Raffo and Lhuillery (2009), we correct all the corrupted characters from the CEMI's PATSTAT[5] Knowledge Base, "Ecole Polytechnique Fédérale de Lausanne" (http://wiki.epfl.ch/patstat/cleaning), and from Lars Tönqvist's typography (http://www.thesauruslex.com/typo/eng/enghtml.htm) for the encoding of foreign characters in HTML. The idea is to replace these types of characters with the corresponding characters in the Latin alphabet which can be easily read by the name matching algorithm. For instance, we make the following changes:

[5] PATSTAT stands for Worldwide Patent Statistical Database.

- 'Ã„' turns into 'AE'
- 'Ã©' turns into 'e'
- 'Ã¶' turns into 'oe'
- 'Ã¼' turns into 'u'
- And so on (see http://wiki.epfl.ch/patstat/cleaning)

Non-HTML-legible foreign characters (vowels with accents, swung dashes, diereses, and so forth) are also modified. A few examples are:

- '&Aacute' is 'Á' and turns into 'A'
- '&Oslash' is 'Ø' and turns into 'O'
- '&aring' is 'å' and turns into 'a'
- '&#274' is '$\overline{E}$' and turns into 'E'
- And so on (see http://www.thesauruslex.com/typo/eng/enghtml.htm)

We also replace all the non-corrupted accentuated characters with their non-accentuated counterparts. The last cleaning-up task is to upper case all the characters and drop slashes, hyphens, accents, diereses, and so on. The full list of changes made is presented in Appendix 2.

Secondly, we harmonize the field as far as possible by placing the surname(s) of the inventor, the first name, and the middle name in different fields. The idea is to use both the surname and the first name as the basis for the subsequent algorithm (see the next subsection).

The middle name may include: the real middle name, or middle names, or initials or other kind of information such as the inventor's affiliation, a surname modifier, and so on. In fact, when surname modifiers or the inventor's affiliation are present, we place them in separate fields and use them as additional information to test whether or not a pair of records belongs to the same inventor. Specifically, we place in a separate field all the information contained in the inventor's name field preceded by 'C/O' as the inventor's potential affiliation.[6] Moreover, we extract an arbitrary list of surname modifiers from this same field and place them in a separate field as well. Examples are 'Prof.', 'Dr.', 'Prof.-Dr.', 'Ing.', 'Jr.', 'PhD.', 'Chem.': for a full list, see Appendix 3.

[6] Other substrings have been used to identify the affiliation of the inventor when placed in the inventor's name field. Some of them are: 'SOCIE', 'GLAX', 'PHILIPS', 'VTT', 'UNIVERSI', 'INTERNATION', 'NATIONAL', or 'INSTITUT'.

For inventor's address, the cleaning-up process resembles the process used for inventor's name, regarding corrupted characters and so on. With regard to the harmonization of fields, we proceed by placing the single address (name of the street and building number), the zip code, and the name of the city in a different field. These three fields will be used in the filtering stage.

Moreover, additional information is retrieved from REGPAT. We also make use of the work carried out by the OECD in this database. Even though PATSTAT users usually have access to country codes linked to inventors' and applicants' patents, it is left to the researcher to find supplementary information at a more refined spatial level regarding the origin of the patent. Additional information can also be found in REGPAT. Maraut et al. (2008) use the address fields of both inventors and applicants of patents to link them to micro-regions in OECD countries. For Europe, the case that interests us here, patents are assigned to NUTS3[7] regions. Basically, the zip codes contained in that field are isolated and linked to the latest version of the NUTS classification code (corresponding to 2006). When the zip code does not appear in the field, the city's name is used instead. From the NUTS3 codes, one can easily retrieve the NUTS2 code for use in the final stage of the present methodology.

The name matching stage

As we said earlier, most of the algorithms found in the literature use the inventor's name and surname to decide "who is who" in the "names game". However, even after cleaning, standardizing, and harmonizing these fields, we may find a string of two inventors' names that actually belong to the same inventor but are assigned to different people – for example, due to spelling errors. Therefore, the second step consists in encoding the strings of the fields mentioned in order to minimize these spelling problems which have introduced variations of the same inventor name. So the name matching algorithm helps us to minimize the ***Type I error***[8].

Name matching algorithms are designed to solve spelling problems like the ones described above. Actually, name variation takes many forms. As reviewed in the literature (Branting, 2003; Snae, 2007) the sources of mistakes may be character variations, including capitalization (Trippl *versus* trippl), punctuation (López Bazo *versus* López-Bazo), spacing (ERNESTMIGUELEZ *versus* ERNEST MIGUELEZ), or qualifiers (Rosina Moreno *versus* Prof. Dr. Rosina Moreno). Some of these sources of problems can be solved through the previous stage. However, other sources of mistakes are spelling variations, including *insertion* (McCann *versus* MacCann),

[7] NUTS stands for the French acronym "*Nomenclature des Unités Territoriales Statistiques*".
[8] The "***Type I error***" occurs if we under-match records, i.e. if we miss records that should be compared to establish whether or not they match, but instead we regard them from the start as different inventors.

omission (Iammarino *versus* Iamarino), *substitution* (Maier *versus* Mayer), or *transposition* (Fingelton *versus* Fingleton). And finally mistakes may arise due to phonetic variations (Cooper in English would be spelled Cuper in German).

A name matching system must deal with cultural as well as spelling and phonetic aspects (Snae, 2007). For instance, there are spelling analysis-based algorithms (like the Guth and Levenshtein algorithms), based on sequences and character strings. There are also phonetics-based algorithms (like Soundex, Metaphone or Phonex), and some composite (ISG) or hybrid (LIG) examples. Given the features of our dataset (with a predominance of English and German-origin names), phonetic algorithms seem to be the most suitable. Among them, the Soundex algorithm is one of the most widely used. Although it was initially designed for English names, it has been extended to other languages. It is the name matching algorithm used in Trajtenberg et al. (2006) and Kim et al. (2006) as well, and, as the authors recognize, the algorithm is quite reliable except for Asian names (whose presence in our dataset, we suspect, will be nominal).

Soundex was developed in the 1930s by the US Census Bureau and was used to list all the individuals in the US census records since 1880. It encodes by using the first letter of each string followed by a number of digits representing the phonetic categories of the next consonants. The vowels and the consonants H, W and Y are ignored, and adjacent letters from the same category are encoded with a single digit. The 0 is used when the string finishes before the whole number of digits has been used. The rest of the letters are encoded as follows:

Table 1. Soundex coding scheme

1	B, P, F, V
2	C, S, K, G, J, Q, X, Z
3	D, T
4	L
5	M, N
6	R

In the present paper, we encode the surname with the first letter of the string and six additional digits, and the name of the inventor using the initial letter and again six additional digits. Combining the Soundex-codes of the surname and the name, we build what Trajtenberg et al. (2006) term *p-sets* (potentially the same inventor). Each different *p-set* is therefore identified as a different, unique inventor. In this way, with the same Soundex-code, we encode the strings that differ slightly but actually belong to the same person (like those of the above examples).

Notwithstanding, this procedure may induce another important error: that is, when two records which actually belong to different inventors are matched under the same *p-set*. Thus, clearly different individuals such as 'Jan Dahlin', 'Jean Pierre Delaunoy', 'Jean Louis Daulon', 'Jean Alain Dalmon', 'Jean Jacques Dulin', 'Joaquim Joao Delima', 'John Lionel Delany' will share the same *p-set* code, D450000J500000 – although, obviously, they are not the same person. Of course, Soundex will encode two researchers named "John Smith" with the same code, even though they do not correspond to the same person. To solve these two types of error, we need to go on to the third stage of the methodology.

The filtering stage

In this third step we perform pair-wise comparisons within each group of possible same inventors in order to minimize ***Type II errors***[9]. The approach chosen in this stage resembles the methodologies used by Lissoni et al. (2006) and Trajtenberg et al. (2006).

We run as many tests as the raw data permit, squeezing all the information linked to each patent in order to optimize the identification procedure. We then assign an arbitrary score to each comparison made and add up the total scores for every pair-wise comparison. This produces the "similarity score" for pairs of inventors with the same Soundex code. We then compare it with a pre-determined numerical threshold, which we use to decide whether two records belong to the same inventor or not. After this, transitivity must be imposed in the sense that, although two inventors, say A and C, are not considered to be the same person – i.e., their "similarity score" derived from their multiple comparisons does not reach the minimum threshold – we impose that they are the same person if A is the same person as B and B is the same as C.

The code to run the pair-wise comparisons was written with Java using the Netbeans software.[10] In Table 2 (section 3) we show the tests we have performed, and the scores assigned to each test. Basically, all the information retrieved is taken from the patent document itself, with few exceptions. As stated above, patent document information is stored in various databases. PATSTAT is the original one, but we use the information stored in the REGPAT database prepared by the OECD; REGPAT contains basically the same information as PATSTAT, but it includes information from the region corresponding to the inventors' addresses reported in the document. The NUTS3 code is therefore included, from which the NUTS2 code can easily be retrieved, if necessary. As far as the applicants are concerned, we use data from the KITES-

[9] "***Type II errors***" are the ones incurred when we end up matching records that in fact belong to different inventors.
[10] Ismael Gómez-Miguélez is the main author of the code.

PatStat database (Bocconi University – Milan). With the applicants' data, the KITES group assign a code to each firm trying to avoid spelling problems and corrupted characters, and also ensuring that an applicant is given the same code even though its applications may be made under different names (for instance, 'I.B.M.' and 'International Business Machines' are assigned the same code). Additionally, KITES gives a group code to each patent if it can be retrieved from 'Dun&Bradstreet'. The idea is that in a few cases, different applicants may belong to the same corporative group, and therefore this information can be used to identify inventors.[11] Citation data to test whether one inventor cites the other one are taken from the 'OECD EP/WO Citation database', which stores citation data that are also contained in patent documents. Here we show the complete list of tests run:

- ***Inventor's bibliographical information***
  - *Same middle name (encoded using Soundex with 6 digits)*
  - *Same inventor's name modifier*
  - *Same affiliation*
  - *Rare pset*
- ***Inventor's bibliographical information from the 'address' field.***
  - *Same street name and building number*
  - *Same zip code*
  - *Same city*
  - *Same NUTS3 region code*
  - *Same NUTS2 region code*
- ***Information from the patent itself: applicant(s) and technological class(es)***
  - *Same applicant code (according to the KITES-PatStat codification)*
  - *Same company code (according to the KITES-PatStat codification)*
  - *Same group code (according to the KITES-PatStat codification)*
  - *Same technological class(es) –IPC code (4 digits)*
  - *Same technological class(es) –IPC code (6 digits)*
  - *Same technological class(es) –IPC code (12 digits)*
- ***Citations information***
  - *If one patent cites the other*

[11] We use the KITES databases due to our participation in the APE-INV project, led by Francesco Lissoni, from the KITES research group. We are very grateful for the opportunity to take part in the project, which in fact enabled us to carry out the present research.

3. Testing the algorithms: The benchmark dataset

Once the three-step methodology is designed, it should be applied to real patent data. The main problem is that we have no way of ascertaining whether the methodology proposed in the present study (as well as other similar methodologies shown elsewhere) is good enough to identify individual inventors. In trying to overcome this difficulty, we use a sample which has been checked manually. Using this benchmark, we decide on a scoring scheme that will give us the highest goodness-of-fit, and we apply this same scoring scheme (and threshold) to the whole dataset. We acknowledge, however, that this procedure is dependent on the "quality" of the benchmark, that is, on the extent to which this benchmark is truly representative of the whole dataset.

The benchmark used is the one designed by Carayol and Cassi (2009), which we were able to access through the APE-INV project. Obviously, we are indebted to them for their invaluable work in manually checking the sample.

The French academic inventors' benchmark

This benchmark comprises 424 French academic inventors (see Lissoni et al., 2010; and Lissoni et al., 2008; for an in-depth description), affiliated to French universities in 2004-2005. This set of inventors is the result of matching EPO patents between 1975 and 2001 with a French ('FR') country code, extracted from the already cleaned KITES-PatStat database, with the list of *'Maitres a Conference'* and *'Professeurs'* listed on French ministerial records in 2005. The total number of patents belonging to each of these academics were also manually checked by Carayol and Cassi (2009) and Lissoni et al. (2010). For our interests, these 424 inventors correspond to 1850 EPO patent applications, and 1996 pairs of Person_IDs and EPO Publication Numbers. However, we use a modified version of this benchmark, which includes additional artificially created homonymy (see Lissoni et al., ibid.). This "noisy" version contains 1950 patent applications and 2097 pairs of Person_ID and EPO publication numbers.

Goodness-of-fit measures and approach used

Before going further, we now show the measures chosen to assess the goodness-of-fit of our algorithm *vis-à-vis* different scoring schemes and thresholds:

The precision rate is:

$$\Pr ecisionRate(PR) = \frac{TruePositives}{TruePositives + FalsePositives}$$

The recall rate is:

$$\mathrm{Re}\, callRate(RR) = \frac{TruePositives}{TruePositives + FalseNegatives}$$

Where:

- **True Positives** are each pair of patents belonging to a given same inventor in the benchmark which the algorithm also identifies as belonging to the same inventor.
- **False Positives** are each pair of patents not belonging to a given same inventor in the benchmark which the algorithm does identify as belonging to the same inventor.
- **False Negatives** are each pair of patents belonging to a given same inventor in the benchmark which the algorithm does not identify as belonging to the same inventor.
- And, for information, **True Negatives** are each couplet of patents not belonging to a given same inventor in the benchmark which the algorithm does not identify as belonging to the same inventor.

We turn now to the description of our approach. As is well known, one of the main problems in this type of exercise is the decision regarding the weights that should be assigned to each of the characteristics tested. Earlier studies have not established a common approach, and some of them give a relatively homogeneous score to each test (Lissoni et al., 2006). Others give different scores that assign an (arbitrary) level of importance to each test (Trajtenberg et al., 2006), whilst some other examples merely decide whether or not two equal names belong to the same person if they share a common, arbitrary characteristic – like the technological class at 4 digits (Agrawal et al., 2006, or other characteristics in the case of Hoisl, 2006, and Kim et al., 2006). A recent study by Carayol and Cassi (2009) is the first attempt to "estimate" the scores and thresholds, giving a "true" sample.

In an attempt to keep things simple, here we start with a homogeneous scoring scheme, as in Lissoni et al. (2006). We give different values to one of the parameters, specifically the threshold up to which a given pair of records is said to belong to the same inventor, and we present the results for 31 different thresholds. We repeat this same procedure using different scoring schemes, by giving heterogeneous scores to the tests, following previous studies (Agrawal et al., 2006; Trajtenberg et al., 2006), and our own common sense. None of these alternative scoring schemes can be said to be superior to the one above (they can be provided upon request from the authors). In table 2 below, we recall the tests applied and show the scores given to each test.

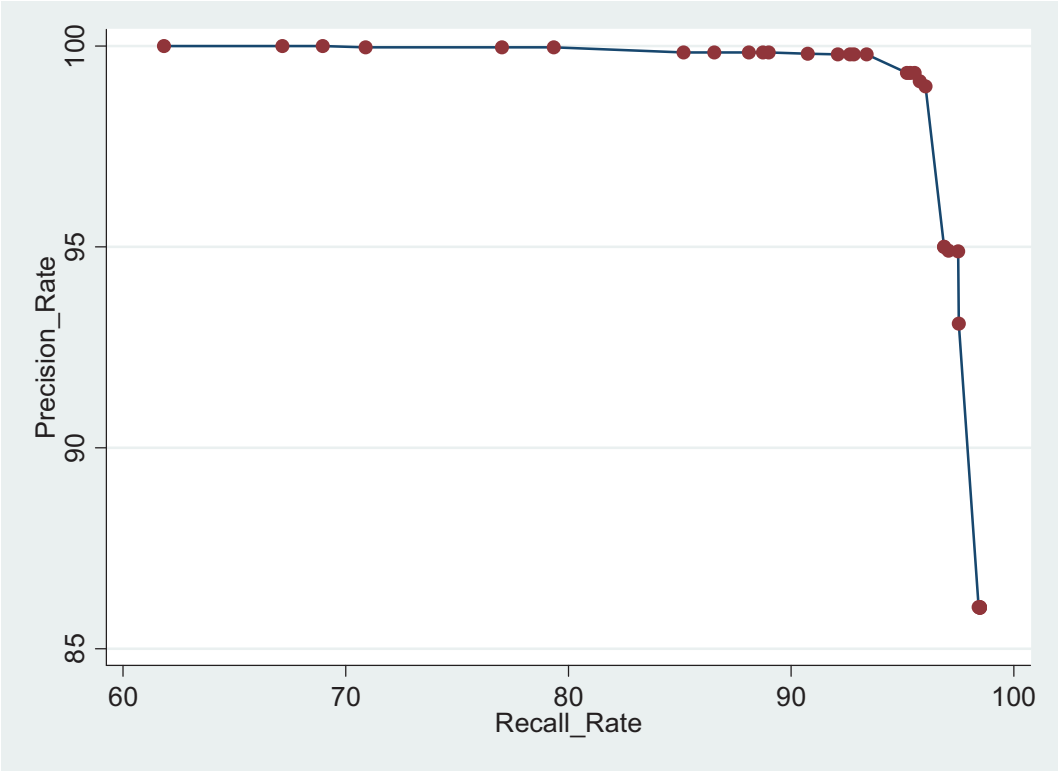
Table 2. Tests and scores of each test

Test	Scores
Same middle name Soundex-code	5
Same surname modifier (if it exists)	5
Same affiliation (if it exists)	5
Rare surname+name Soundex-code	5
Same street and building number	5
Same ZIP code	5
Same city	5
Same NUTS-3 region	5
Same NUTS-2 region	5
Same applicant code	5
Same company code (if it exists)	5
Same group code (if it exists)	5
Same technological class (4 digits)	5
Same technological class (6 digits)	5
Same technological class (12 digits)	5
Self-citation	5

Results on the French academic inventors' benchmark

In Figure 1 and Table 3 we show the results of the algorithm applied to the French noisy benchmark, using the scoring scheme detailed in Table 2 and different thresholds, from 0 to 30. As can be seen, the precision and recall rates are very high. We can also choose the threshold that best suits our purposes. Figure 1 below shows the points resulting from the combination of recall and precision rates.

Figure 1. Goodness-of-fit: recall and precision rates



Given that the main purpose of the subsequent econometric estimations is the study of inventors' professional and geographical mobility and of the strength and scope of their collaboration networks, we are especially interested in minimizing the number of false positives (that is each pair of patents which do not belong to the same inventor in the benchmark but which the algorithm identifies as belonging to the same inventor) but without compromising the number of false negatives. Consequently, given the scoring scheme mentioned above, by setting the threshold at 15 we have a reasonably limited number of false positives (32) and the lowest number of false negatives among the thresholds with only 32 false positives. Note that when the threshold rises from 14 to 15, the number of false positives falls sharply, while going beyond 15 the number does not fall substantially, while the number of false negative increases steadily.

Table 3. Results with the French benchmark for different thresholds

True Positives	True Negatives	False Positives	False Negatives	Threshold	Precision Rate	Recall Rate
17,180	4,375,074	2,792	266	0	86.02	98.48
17,180	4,375,074	2,792	266	1	86.02	98.48
17,180	4,375,076	2,790	266	2	86.03	98.48
17,180	4,375,076	2,790	266	3	86.03	98.48
17,174	4,375,078	2,788	272	4	86.03	98.44
17,018	4,376,604	1,262	428	5	93.10	97.55
17,010	4,376,950	916	436	6	94.89	97.50
16,938	4,376,958	908	508	7	94.91	97.09
16,902	4,376,976	890	544	8	95.00	96.88
16,902	4,376,976	890	544	9	95.00	96.88
16,758	4,377,696	170	688	10	99.00	96.06
16,712	4,377,720	146	734	11	99.13	95.79
16,672	4,377,754	112	774	12	99.33	95.56
16,636	4,377,756	110	810	13	99.34	95.36
16,610	4,377,756	110	836	14	99.34	95.21
16,294	4,377,834	32	1,152	15	99.80	93.40
16,194	4,377,834	32	1,252	16	99.80	92.82
16,162	4,377,834	32	1,284	17	99.80	92.64
16,068	4,377,834	32	1,378	18	99.80	92.10
15,834	4,377,836	30	1,612	19	99.81	90.76
15,528	4,377,842	24	1,918	20	99.85	89.01
15,482	4,377,842	24	1,964	21	99.85	88.74
15,372	4,377,842	24	2,074	22	99.84	88.11
15,100	4,377,842	24	2,346	23	99.84	86.55
14,858	4,377,844	22	2,588	24	99.85	85.17
13,842	4,377,862	4	3,604	25	99.97	79.34
13,436	4,377,862	4	4,010	26	99.97	77.01
12,370	4,377,864	2	5,076	27	99.98	70.90
12,032	4,377,866	0	5,414	28	100.00	68.97
11,716	4,377,866	0	5,730	29	100.00	67.16
10,786	4,377,866	0	6,660	30	100.00	61.83

4. Whole patent dataset and descriptive statistics

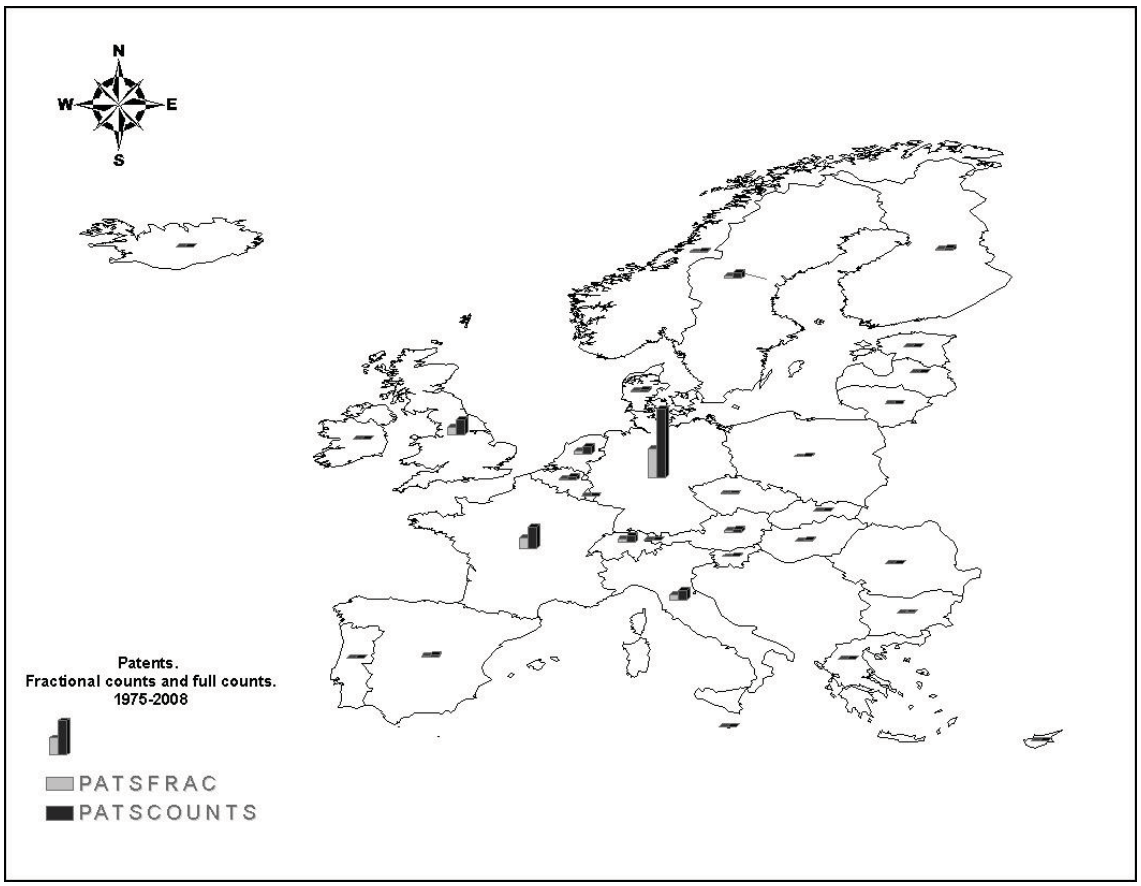
In this section, we apply the methodology described so far to the whole dataset of patents. Specifically, we apply the procedures to the REGPAT database (OECD, January 2010 edition). First we briefly describe the data used, alongside a number of figures. Then we present a summary of results in terms of inventors identified, their average characteristics, their technological and spatial distribution, and their evolution over time.

The REGPAT database for Europe

The raw data for our study were collected from the OECD REGPAT database (OECD, January 2010 edition). This dataset uses data from the PATSTAT database to link the addresses of the inventors and applicants of each patent to more than 2,000 regions throughout the OECD countries (see Maraut et al. (2008) for a description of the methodology). Thanks to their fruitful work, we can identify the region in which each inventor works when she applies for a patent. Basically, they focus on the process of regionalization of patent data at very low levels of disaggregation, which they assess using the addresses of the inventor recorded in patent documents (the ZIP code or, in its absence, the town name). This regionalization procedure provides researchers with a complete dataset of patents applied for at the European Patent Office, and contains a wealth of information, i.e., the publication number, the priority year (that is to say, the year when a patent was filed for the first time), information on the name, address, region code and country code of the inventor(s) and applicant(s) of each patent, the share of the patent that corresponds to each inventor or applicant – in order to account for co-authorships and multi-applicants – and finally the technological class(es) to which each patent corresponds.

We restrict our identification methodology to inventors living in European countries. The full list of countries is shown in Appendix 4. From a time dimension perspective, we exploit all the data available and hence have data from 1978 to 2005. According to Maraut et al. (2008), the OECD's regionalization process reached a success rate of 98% for the case of EPO patents. However, for some countries this process ended up in allocations of NUTS codes with a breakdown – for the case of Germany, for instance, the share of addresses with a breakdown into different NUTS3 is around 14% (Ibid.). Since correct regionalization is a priority for us in order to be able to study mobility across regions, we remove all the patents with a regionalization breakdown below 70%. Additionally, for some addresses no allocation is obtained, for various reasons: town names allocated to different NUTS3 regions, addresses referring to a wrong country, the address field is empty or not valid, and so on. We also remove all these patents. All in all, however, the number of records eliminated for these reasons does not exceed 1.8%. Our final dataset contains 2,297,196 records, corresponding to all the pair-wise combinations of inventors' name strings plus the patent number, from 1978 to 2005. This corresponds to 1,041,080 different patents, representing an average number of different inventors per patent of around 2.21. The distribution of EPO patents across countries is highly unbalanced (see Figure 2); Germany is the most productive country in terms of innovation outputs, followed by France and Great Britain, regardless of whether patents are aggregated by fractional or full counts. The last country in terms of patents production is Malta.

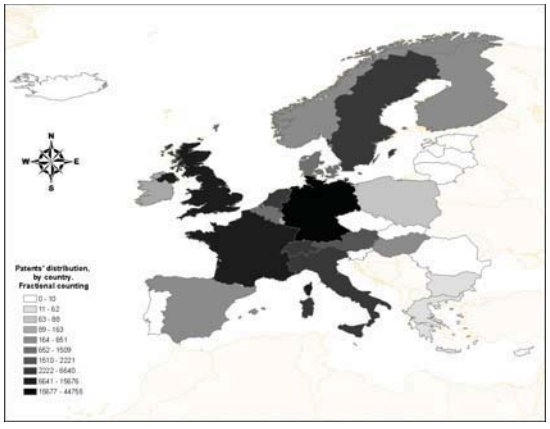
Figure 2. Distribution of patents across European countries, fractional counts and full counts. 1978-2005



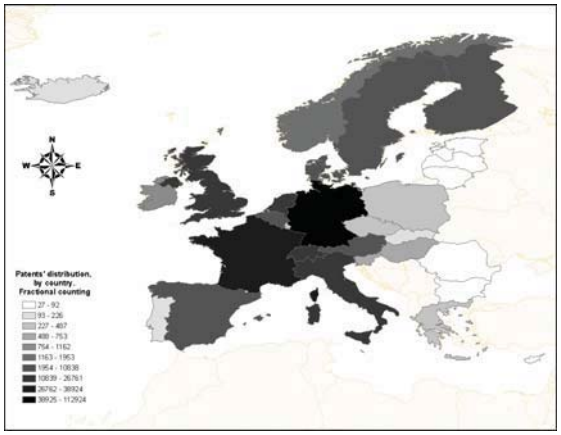
Additionally, this uneven distribution remains practically unchanged over time. Figure 3 shows the distribution of patents across countries at two different points in time, separated by a 20-year gap.

Figure 3. Distribution of patents across countries, fractional counts.

i) 1981-1985

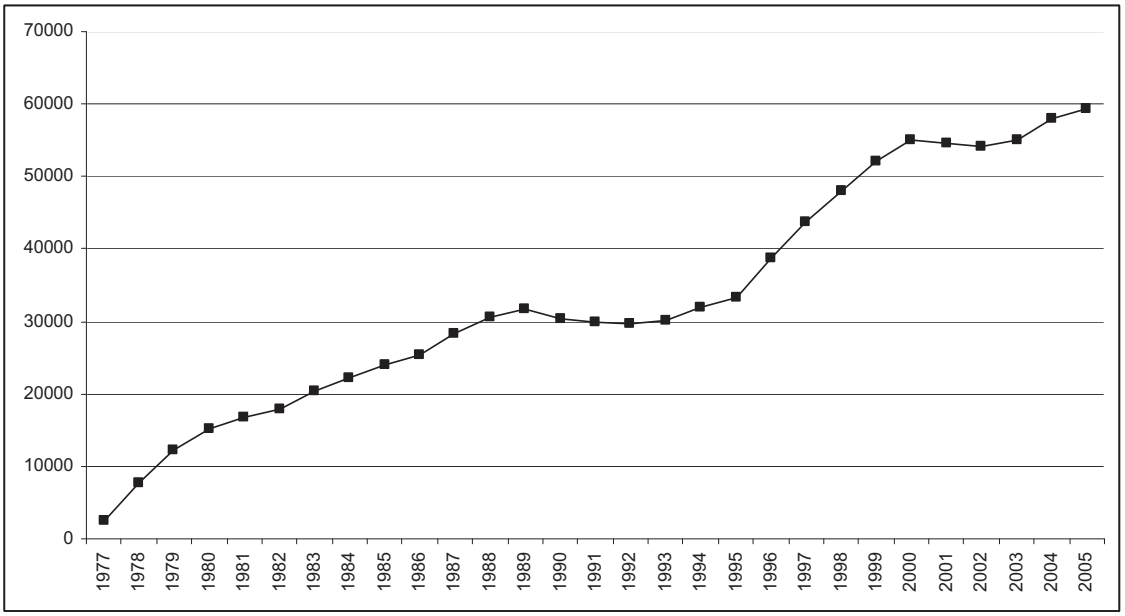


ii) 2001-2005



The patent activity in the sample shows a continuous upward trend in the number of applications throughout the period. Among the few exceptions are the period of recession in the early nineties, and a small stagnation in the production of patents between 2001 and 2002, coinciding with the "dot-com bubble". In any case, the overwhelming general increase in patent production can be attributed to the rising technological complexity of economic activity, as well as the increase in the use of the European Patent Office either instead of, or in complement to, national offices.

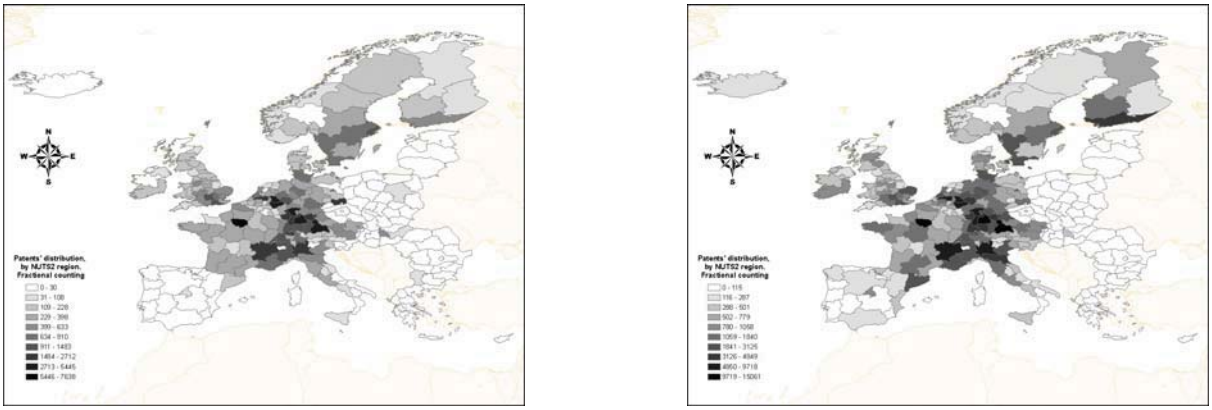
Figure 4. Patent evolution in Europe: fractional counts. 1977-2005



The spatial distribution of patents is even more unbalanced if we look at the regional level (NUTS2 level of regional desegregation). The two maps in figure 5 correspond to the regional distribution of patents at separate moments in time. As we can see, this distribution is very uneven as well, and in some cases it is also uneven inside countries – in the UK and Spain, for instance. Regarding the time dimension, more regions show dark shades in the second period than in the first one, though the differences in patent production remain large and virtually unchanged across time for the majority of regions.

Figure 5. Distribution of patents across NUTS2 regions, fractional counts.

i) 1981-1985 **ii) 2001-2005**



Results of the different stages of the methodology

The parsing stage

After the parsing stage – cleaning, harmonizing and standardizing the inventor's name field and the address field – a few figures stand out. For instance, the initial 2,297,196 records are made up of 29,017 different names, 257,227 surnames, and 678,324 combinations of names and surnames. Additionally, 509,597 of the 2,297,196 records (22.18%) have a middle name (or the initial). In 300,523 cases (13.08%) there is a surname modifier, and in 30,262 records (1.32%), the affiliation of the inventor can be retrieved. The following table presents the most common names, surnames, and combinations of both.

Table 4. Top ten frequency of names, surnames, and name-surname.

Name	# records	Surname	# records	Name+Surname	# records
PETER	50,058	MULLER	10,758	EBERHARD AMMERMANN	526
JEAN	48,213	SCHMIDT	7,289	VOLKER REIFFENRATH	481
HANS	47,832	FISCHER	5,210	ROBERT SCHMIDT	473
MICHAEL	37,625	SCHNEIDER	4,761	HEINZ FOCKE	446
THOMAS	33,710	WEBER	3,825	HANS SANTEL	406
WOLFGANG	29,232	MEYER	3,586	GISELA LORENZ	381
KLAUS	28,673	BAUER	3,142	KLAUS MULLER	377
MARTIN	22,362	WAGNER	3,058	HANS MULLER	346
KARL	21,218	MARTIN	2,838	JEAN GUERET	344
ANDREAS	20,753	SMITH	2,792	SIEGFRIED STRATHMANN	340

As for the addresses, the records are distributed in 127,131 different zip codes, 151,582 cities and towns, 1,312 NUTS3 regions, and 289 NUTS2 regions. Table 5 shows the most repeated zip codes, cities, NUTS3 and NUTS2 in terms of numbers of records.

Table 5. Top ten frequency of zip codes, cities, NUTS3 and NUTS2.

Zip code	# records	City	# records	NUTS3	# records	NUTS2	# records
5656	40,019	MUNCHEN	43597	NL414	49,120	FR10	136,638
8000	20,003	EINDHOVEN	35531	FR101	38,356	DE21	105,090
8501	7,478	PARIS	33611	DE212	35,132	DE11	97,669
1000	7,456	BERLIN	26881	ITC45	30,364	DE71	92,653
5000	6,605	STUTTGART	15004	FR105	28,974	DEA1	85,845
5090	6,590	HAMBURG	13622	DE300	27,107	DEA2	76,701
5600	5,630	KOELN	13362	SE110	24,703	DEB3	67,021
6700	5,501	LEVERKUSEN	11537	CH040	23,873	DE12	59,475
4000	5,157	MILANO	11446	DE115	22,628	NL41	57,010
75008	5,139	DUSSELDORF	11334	FR103	20,648	FR71	52,932

The matching stage

After applying the name matching algorithm, that is, the Soundex code for names and surnames, several points should be stressed. Recall from the previous sections that this algorithm avoids the spelling problems that introduce variation in the inventors' name field if a given pair of records belongs to the same inventor. Unfortunately, however, this algorithm forces us to compare two clearly distinct names that may share the Soundex code for name and surname. As a result of applying the name matching algorithm, we ended up with 379,030 different Soundex codes. In Table 6 below, the most repeated codes are shown, alongside their frequency within our dataset. Thus, on average, every different Soundex code comprises 1.79 clearly different combinations of name and surname – which, however, may be due to completely different names, or due to misspellings of the same name. Table 6 includes a few examples of both situations for the case of the most frequent Soundex code. On average, every Soundex code contains 6.06 records.

Table 6. Top ten frequency of Soundex codes and ten examples of the first.

Soundex code pset	# records		Most freq. pset	Surname, name and middle name
M460000H520000	887		M460000H520000	MULLER, HENNING
M600000J500000	660		M460000H520000	MULLER, HEINZ K
G630000J500000	654		M460000H520000	MULLER, HEINZ KONRAD
M200000J500000	651		M460000H520000	MULLER, HANS WILLI
R200000J500000	646		M460000H520000	MULLER, HANNS PETER
S530000R163000	605		M460000H520000	MOELLER, HENNING
F200000H520000	601		M460000H520000	MOELLER, HENNING BIRGER
B200000J500000	587		M460000H520000	MEILER, HANS ECKHARD KAUFMANN
S530000H520000	579		M460000H520000	MEILER, HANS ECKHARD KFM
S530000J500000	564		M460000H520000	MAHLER, HANNS CHRISTIAN

The filtering stage

Applying the three stages using patent data from OECD REGPAT databases (January 2010 edition) we finally identify 768,810 inventors from a sample of 2,297,196 initial records. This means an average of 2.99 patents per inventor, a rate similar to that reported in other studies (see, for instance, Trajtenberg et al., 2006). As Table 7 shows, the distribution of the number of patents per inventor is highly skewed, since the majority of inventors (55.99%) have only one patent and 88.69% have fewer than six. Only 0.23% of the inventors identified have more than 50 patents.

Table 7. Distribution of patents per inventor.

Patents per inventor	**Number of inventors**	**% of inventors**
1	**430,458**	**55.99**
2-5	**251,428**	**32.70**
6-9	**45,579**	**5.93**
10-50	**39,619**	**5.15**
+50	**1,726**	**0.23**
	768,810	**100**

The distribution of the inventors identified across countries is also very uneven. As expected, Germany is the country with the highest number of inventors (as in the case of patents), followed by France and the UK (Table 8 and Figure 6).[12] At the other end of the scale, Malta is the country with the fewest inventors throughout the period.

Table 8. Distribution inventors across countries.

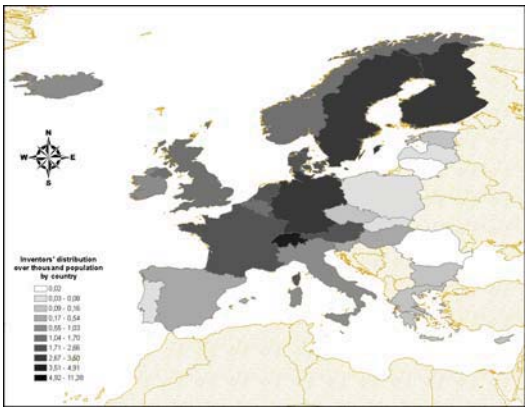
Country name	# inventors	Country name	# inventors
Germany	283,569	Czech Republic	1,646
France	123,829	Greece	1,312
United Kingdom	97,930	Slovenia	1,032
Italy	54,090	Luxemburg	995
The Netherlands	43,399	Bulgaria	820
Switzerland	36,506	Portugal	719
Sweden	31,563	Slovakia	424
Austria	17,897	Liechtenstein	396
Belgium	17,786	Romania	382
Spain	16,236	Iceland	307
Finland	14,910	Estonia	187
Denmark	12,135	Latvia	170
Norway	6,470	Cyprus	107
Hungary	5,397	Lithuania	75
Ireland	3,982	Malta	54
Poland	1,800		

Thus, this unbalanced spatial distribution of inventors is further confirmed in the following maps (Figure 6) where the distribution of inventors over population is depicted both at country level (i) and at the NUTS2 level (ii).

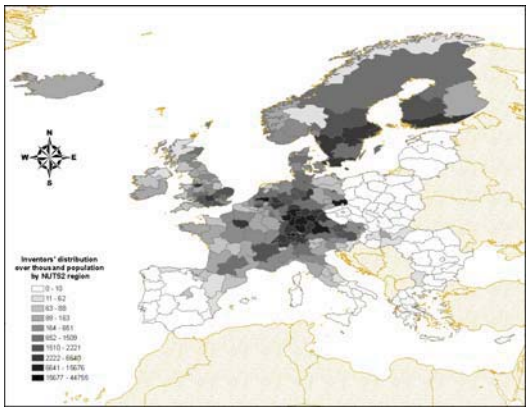
[12] In this general enumeration of inventors across European countries, we ignore the possibility of migration. Thus, if an inventor appears in two distinct countries or regions, he/she is counted twice.

Figure 6. Distribution of inventors over population across countries and NUTS2 regions

i) NUTS0



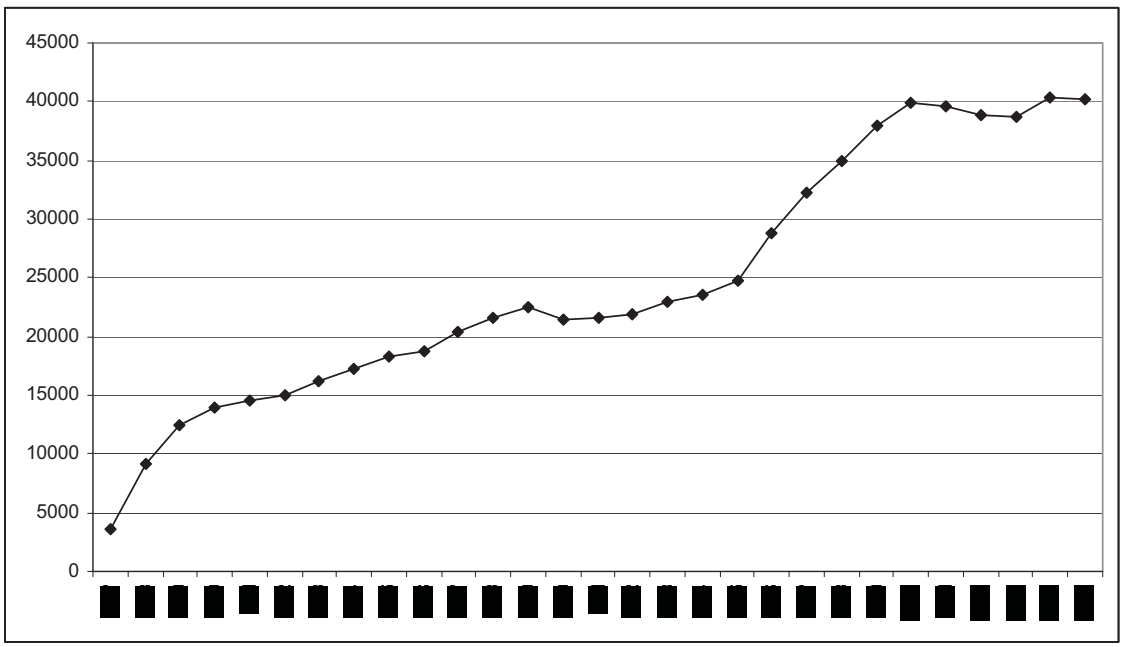
ii) NUTS2



Note: To calculate this ratio, we compute all the inventors identified throughout the period of analysis over population in 2005.

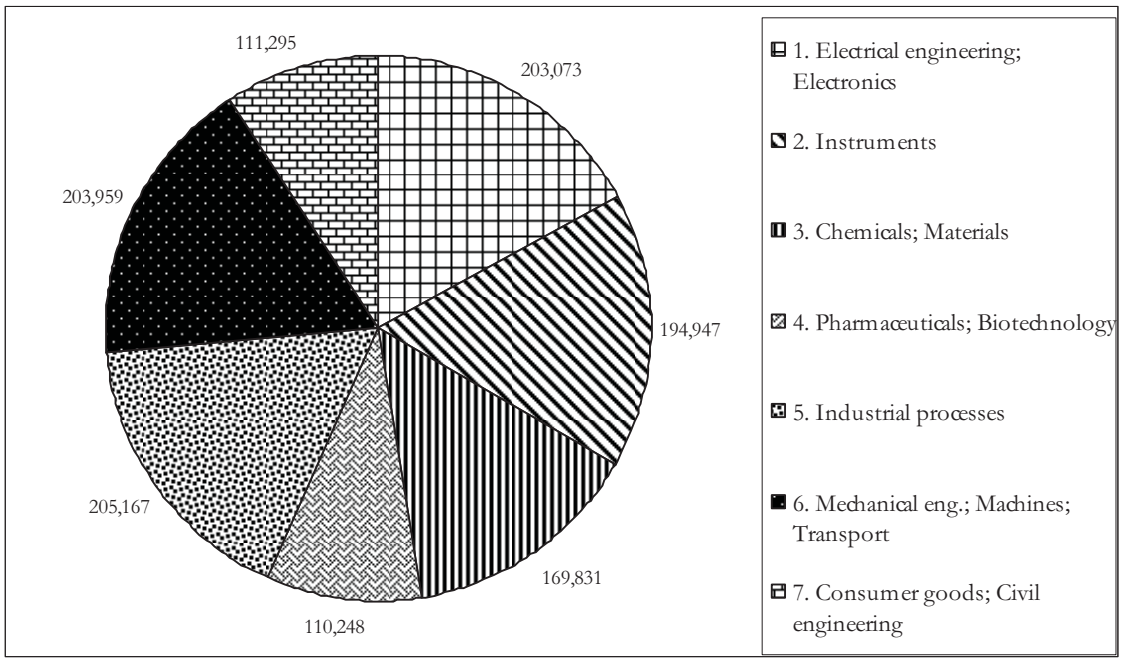
Figure 7 below shows the evolution over time of the level of inventors in Europe. The allocation of inventors in time is done using the priority date of their first application. Obviously, both the spatial distribution of inventors and their time evolution is highly dependent upon the number of patents applied for at the EPO. At the same time, however, spatial distribution and evolution over time of patent applications are highly dependent upon the presence/existence of inventors in given locations and time periods, so the descriptive analysis of inventors' distribution in space and time is interesting in itself.

Figure 7. Inventors' evolution in Europe. 1977-2005



Another interesting point is the distribution of inventors across technological sectors.[13] Figure 8 below shows this distribution across technologies for the whole period under analysis (1977-2005). As can be seen, industrial processes, mechanical engineering, and electrical engineering are the sectors with the most inventors. However, in contrast to their spatial distribution, the differences across technological sectors are not that pronounced.

Figure 8. Inventors' distribution across technological sectors. 1977-2005



The following figures (Figures 9 and 10) also show the evolution of inventors in time across different sectors. In spite of the increase in the quantity of inventors in all seven sectors, their relative importance has changed slightly during the whole period. Although their respective share remains stable through time (Figure 10), several changes may be reported: sectors like electrical engineering and pharmaceuticals and biotechnology have increased in importance, whilst

[13] As regards the technological classification used to describe the distribution of inventors across technological sectors, we adopt a technology-oriented classification designed jointly by Fraunhofer Gesellschaft-ISI (Karlsruhe), Institut National de la Propriété Industrielle (INPI, Paris) and Observatoire des Sciences and des Techniques (OST, Paris). This classification aggregates all IPC codes into seven technology fields: 1. Electrical engineering; Electronics (including Electrical engineering, Audiovisual technology, Telecommunications, Information technology, Semiconductors); 2. Instruments (including Optics, Technologies for Control/Measures/Analysis, Medical engineering, Nuclear technology); 3. Chemicals; Materials (including Organic chemistry, Macromolecular chemistry, Basic chemistry, Surface technology, Materials; Metallurgy); 4. Pharmaceuticals; Biotechnology (including Biotechnologies, Pharmaceuticals; Cosmetics, Agricultural and food products); 5. Industrial processes (Mechanical engineering (excl. Transport), Handling; Printing, Agricultural and food apparatuses, Materials processing, Environmental technologies); 6. Mechanical eng.; Machines; Transport (Machine tools, Engines; Pumps; Turbines, Thermal processes, Mechanical elements, Transport technology, Space technology; Weapons); and 7. Consumer goods; Civil engineering.

industrial processes have fallen off. However, the number of inventors has increased sharply in all the sectors.

Figure 9. Inventors' evolution by technological sector. 1977-2005

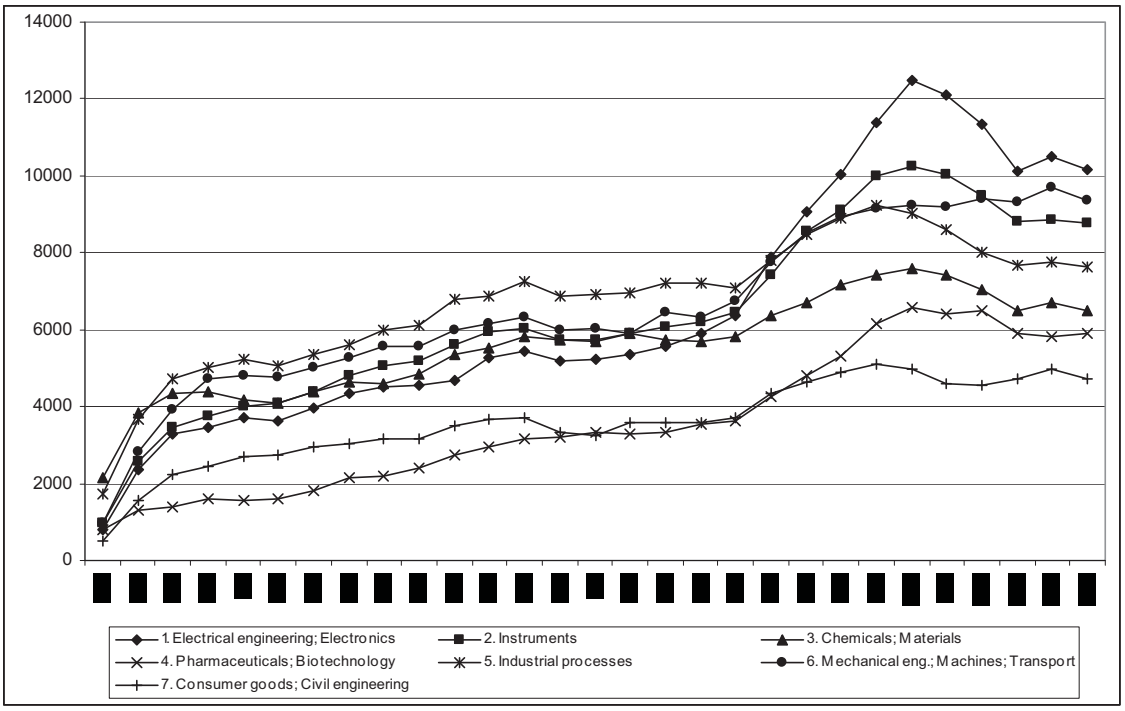
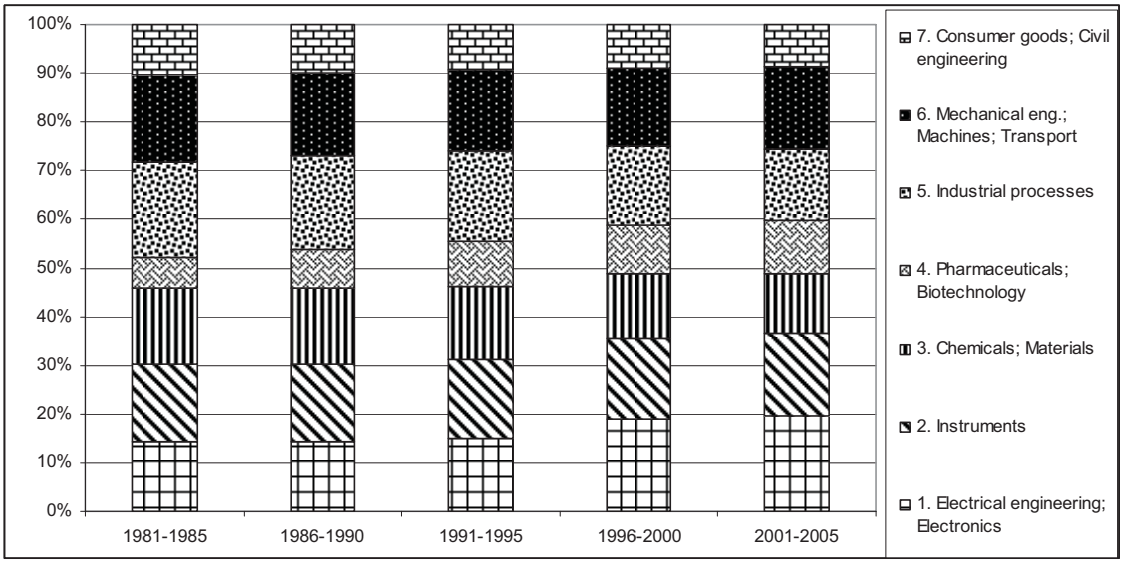


Figure 10. Inventors' distribution across technological sectors and time periods. 1977-2005



5. Conclusions

We describe in detail our methodology for identifying individual inventors through the use of patent documents. To recap, this methodology comprises three steps: first, a cleaning-up process of the raw data; second, the use of SOUNDEX, a name matching algorithm, in order to group possible similar names; and third, a "splitting" algorithm to ascertain whether pairs of grouped inventors are the same person or not. To undertake this final step we suggest a set of tests which use as much information as possible from the patent document itself. We assign a score to each test and then add up the scores. If the total score reach a minimum threshold, a given pair of inventors were said to be the same person. In order to choose the scores we run our algorithm iteratively for a small sample of French academic inventors for whom we know exactly "who is who". We calculate recall and precision rates (false positives and false negatives) from this benchmark, and use the scoring scheme and threshold which best suits our purposes.

Our procedure for choosing the scores could be criticized, as we were not able to run all the possible combinations of scores and thresholds using all the tests performed. In future research we plan to design an algorithm capable to decide endogenously the scores of the splitting algorithm by itself (this is done in a way by Carayol and Cassi, 2009).

References

Agrawal A, Cockburn I, McHale J (2006) Gone but not forgotten: labour flows, knowledge spillovers, and enduring social capital. Journal of Economic Geography 6: 571-591.

Bottazzi L. and Peri G. (2003) Innovation and spillovers in regions: Evidence from European patent data, European Economic Review 47, 687 – 710.

Branting LK (2003) A comparative evaluation of name-matching algorithms, International Conference on Artificial Intelligence and Law.

Carayol N., Cassi L. (2009) "Who's Who in Patents. A Bayesian approach", Cahiers du GREThA 2009□07, Groupe de Recherche en Economie Théorique et Appliquée – Université Bordeaux 4, Bordeaux.

Fleming, L. and C. King, A. Juda, "Small Worlds and Regional Innovation." Organization Science, Vol. 18, No. 2 (2007), pp. 938-954.

Giuri P, Mariani M, Brusoni S, Grespi G, Francoz D, Gambardella A, Garcia-Fontes W, Geuna A, Gonzales R, Harhoff D, Hoisl K, Le Bas C, Luzzi A, Magazzini L, Nesta L, Nomaler Ö, Palomeras N, Patel P, Romanelli M, Verspagen B (2007) Inventors and invention processes in Europe: Results from the PatVal-EU survey. Research Policy 36: 1107-1127.

Griliches, Z. (1991) "Patent Statistics as Economic Indicators: A Survey," NBER Working Papers 3301.

Hoisl, Karin (2006): German PatVal Inventors – Report on Name and Address-Matching Procedure, unpublished manuscript, University of Munich. http://www.inno-tec.bwl.uni-muenchen.de/files/forschung/publikationen/hoisl/patval_matching.pdf.

Kim J, Lee SJ, Marschke G (2006) International knowledge flows: Evidence from an inventor-firm matched dataset. NBER Working Paper 12692.

Lai R., D'Amour A., Fleming L. (2009) "The careers and co□authorship networks of U.S. patentholders, since 1975", Harvard Business School □ Harvard Institute for Quantitative Social Science.

Lissoni F, Sanditov B, Tarasconi G (2006) The Keins database on academic inventors: methodology and contents CESPRI Working Paper, 181.

Lissoni F, Maurino A, Pezzoni M, Tarasconi G (2010) APE-INV's "name game" algorithm challenge: a guideline for benchmark data analysis & reporting http://www.esf-ape-inv.eu/download/Benchmark_document.pdf.

Maraut S, Dernis H, Webb C, Spiezia V, Guellec D (2008) The OECD REGPAT Database: A presentation STI Working Paper 2008/2.

Raffo J, Lhuillery S (2009) How to play the "Names Game": Patent retrieval comparing different heuristics, Research Policy, In Press: doi:10.1016/j.respol.2009.08.001.

Snae C (2007) A comparison and analysis of name matching algorithms. Proceedings of World Academy of Science, Engineering and Technology 21: 252-257.

Thoma G. and Torrisi S. (2007), Creating Powerful Indicators for Innovation Studies with Approximate Matching Algorithms. A test based on PATSTAT and Amadeus databases.

Trajtenberg M, Shiff G, Melamed R (2006) The "names game": harnessing inventors' patent data for economic research, NBER working paper 12479.

Appendix.

Appendix 1. Compilation of studies aimed to identify individual inventors

Authors, year	Data source	Main methods
Agrawal, Cockburn, McHale (2006)	USPTO data until 1990	▪ Unknown parsing ▪ Exact matching of surname and name ▪ Coincidence of technological class at 4 digits
Carayol and Cassi (2009)	EPO patents with at least one inventor declaring a metropolitan French address, 1977-2003: Additionally, 455 French scholars manually verified.	▪ Standard parsing ▪ No matching algorithm. Spelling problems assumed inexistent. ▪ Bayesian estimation of scores and threshold to minimize precision and recall rates, using information about same first name & name, same assignee, same city, same IPC (6 digits), citation links between pairs of patents.
Hoisl (2006)	EPO (1975-2002) German patents included in the PatVal database	▪ Parsing of corrupted characters and non-Latin characters, removal of accents and use of lower case, split of name, surname, and middle name ▪ Exact matching of last name ▪ The more the conditions met, the higher the probability of correct matching. Conditions: last name, first name, partial first name, street, city, partial city, IPC main, applicant.
Kim, Lee, Marschke (2005)	USPTO, 1969-2002	▪ Unknown parsing ▪ Soundex code of surname and name ▪ One of the following conditions are met: (1) coincidence in full address, (2) self-citation, (3) coincidence of co-inventors
Lai, D'Amour, Fleming (2009)	NBER patent dataset 1975-1999, and USPTO till now	▪ Standard parsing ▪ Matching algorithm: approximate matching, Jaro-Winkler method. ▪ Own algorithm: "adjacency matching": Optimisation of the weights to assign to each comparison. Information compared: name information, assignee information, location information, technology class and co-author data. Inclusion of frequency adjustments
Lissoni, Sanditov, Tarasconi (2006)	EP-CESPRI database, for Italy, Sweden and France	▪ Paring: Elimination of non-letter characters, symbols, accents, ASO. Capitalisation ▪ Same name and surname, exact matching ▪ If equal name+surname but different address, several tests are performed. With almost equal scoring, tests are related to: technological classes, inventors' location, assignee, information about co-authors, cross-citations. Threshold about the mean similarity score.
Raffo and Lhuillery (2009)	Set of inventors applying to EPO affiliated to the *Ecole Polytechnique Fédérale de Lausanne*	▪ Test of various parsing techniques. Better results with additional parsing techniques ▪ Various matching techniques tested. The weighted 2-gram method is found to be the best ▪ Multiple filters using typical information available. Test of optimal threshold.
Trajtenberg, Shiff, Melamed (2006)	NBER patents and citations data file, USPTO patents 1963-1999. The Israeli set of inventors as benchmark	▪ Parsing by eliminating non-letter characters and symbols from the name string, drop blank spaces, and capitalisation ▪ Soundex code of surname and name ▪ Different arbitrary scores given to a set of characteristics tested (in order of importance): full address, self citation, same collaborators, middle name and surname modifiers, assignee, city and technological class of the patent. Arbitrary threshold.

Appendix 2.

Corrupted characters:

'Ã¬'→' '
'Âº'→' '
'Ã□'→'A'
'Ã¡'→'a'
'Ã '→'a'
'Ã¢'→'a'
'Ã„'→'AE'
'Â«'→'AE'
'Ã¤'→'ae'
'Ã£'→'a'
'Ã¥'→'a'
'Ã•'→'a'
'Ã…'→'A'
'Ã¦'→'ae'
'Âµ'→'ae'
'Ã—'→'C'
'Ã§'→'c'
'?'→'E'
'Ã±'→'E'
'Ã©'→'e'
'Ã¨'→'e'
'Ãª'→'e'
'Ã‹'→'E'
'Ã«'→'e'
'Â¢'→'e'
'Ã-'→'i'
'Ã®'→'i'
'Ã¯'→'i'
'Â¾'→'o'
'Ã³'→'o'
'Ã²'→'o'
'Ã´'→'o'
'Ã□'→'OE'
'Ã–'→'OE'
'Ã¶'→'oe'
'Ã·'→'oe'
'Ã”'→'O'
'Ã¸'→'o'
'Ã˜'→'O'
'Ã“'→'O'
'ÃŸ'→'ss'
'Â·'→'u'
'Ãº'→'u'
'Ã»'→'u'
'Â¨'→'U'
'Â©'→'U'
'Ã¼'→'u'
'Â³'→'u'
'Ãœ'→'U'
'Ã¿'→'y'
'→Â¹'→' '
'→Â¹'→' '
'Â-'→'E'
'Ã□'→"
'Â '→"
'Â¿'→"
'Ã'→'N'
'Ã,'→'A'
'Â±'→"
'Â¤'→"
'Â§'→' '
'Â¬'→"
'Ã°'→"
'Ãµ'→'o'
'Ã‰'→"
'Â¼'→"
'Â½'→'A'
'Ã½'→"
'Â¹'→' '
'Ãž'→' '
'Ã□'→'o'
'Â´'→"
'Â®'→'o'
'Â°'→'o'
'Ã¹'→"
'Â²'→'O'
'Ãš'→'e'

Foreign characters:

'Ç'→'C'
'ç'→'c'
'Ë'→'E'
'ë'→'e'
'À'→'A'
'à'→'a'
'È'→'E'
'è'→'e'
'É'→'E'
'É'→'e'
'Í'→'I'
'Í'→'i'
'Ï'→'I'
'ï'→'i'
'Ò'→'O'
'ò'→'o'
'Ó'→'O'
'ó'→'o'
'Ú'→'U'
'ú'→'u'
'Ü'→'U'
'ü'→'u'
'·'→"
'Ć'→'C'
'ć'→'c'
'Č'→'C'
'č'→'c'
'Đ'→'D'
'đ'→'d'
'Š'→'S'
'š'→'s'
'Ž'→'Z'
'ž□'→'z'
'Ď'→'D'
'ď'→'d'
'Ě'→'E'
'ě'→'e'
'Ň'→'N'
'ň'→'n'
'Ř'→'R'
'ř'→'r'
'Š'→'S'
'š'→'s'
'Ť'→'T'
'ť'→'t'
'Ů'→'U'
'ů'→'u'
'Ý'→'Y'
'ý'→'y'
'Æ'→'AE'
'æ'→'ae'
'Ø'→'O'
'ø'→'o'
'Å'→'A'
'å'→'a'
'Ä'→'A'
'ä'→'a'
'Ö'→'O'
'ö'→'o'
'Õ'→'O'
'õ'→'o'
'Ð'→'D'
'ð'→'d'
'Â'→'A'
'â'→'a'
'Ê'→'E'
'ê'→'e'

'Î'→'I'
'î'→'i'
'Ô'→'O'
'ô'→'o'
'Œ'→'OE'
'œ'→'oe'
'Û'→'U'
'û'→'u'
'Ÿ'→'Y'
'Ź'→'y'
'ß'→'B'
'Ő'→'O'
'ő'→'o'
'Ű'→'U'
'ű'→'u'
'Þ'→'P'
'þ'→'p'
'Ā'→'A'
'ā'→'a'
'Ē'→'E'
'ē'→'e'
'Ģ'→'G'
'ģ'→'g'
'Ī'→'I'
'ī'→'i'
'Ķ'→'K'
'ķ'→'k'
'Ļ'→'L'
'ļ'→'l'
'Ņ'→'N'
'ņ'→'n'
'Ŗ'→'R'
'ŗ'→'r'
'Š'→'S'
'š'→'s'
'Ū'→'U'
'ū'→'u'
'Ą'→'A'
'ą'→'a'
'Ć'→'C'
'ć'→'c'
'Ł'→'L'
'ł'→'l'
'Ń'→'N'
'ń'→'n'
'Ś'→'S'
'ś'→'s'
'Ź'→'Z'
'ź'→'z'
'Ż'→'Z'
'ż'→'z'
'Ã'→'A'
'ã'→'a'
'ª'→'a'
'º'→'o'
'Ă'→'A'
'ă'→'a'
'Ş'→'S'
'ş'→'s'
'Ţ'→'T'
'ţ'→'t'
'¡'→''
'¿'→''
'€'→''
'&pound'→''
'«'→''
'»'→''
'•'→''
'†'→''
'©'→''
'®'→''
'°'→''
'µ'→''
'·'→''
'–'→''
'&mdash'→''
'№'→''
'Č'→'C'
'č'→'c'
'Š'→'S'
'š'→'s'

Accents, slashes, diaeresis, and other punctuation symbols:

'Ä'→'A'
'Ë'→'E'
'Ï'→'I'
'Ö'→'O'
'Ü'→'U'
'À'→'A'
'È'→'E'
'Ì'→'I'
'Ò'→'O'
'Ù'→'U'
'Á'→'A'
'É'→'E'
'Í'→'I'
'Ó'→'O'
'Ú'→'U'
'Â'→'A'
'Ê'→'E'
'Î'→'I'
'Ô'→'O'
'Û'→'U'
'Î'→'I'
'{'→' '
'}'→' '
'('→' '
')'→' '
'Ç'→'C'
'Å'→'A'
'Å'→'A'
'Ø'→'O'
'Æ'→'AE'
'Ã'→'A'
'Õ'→'O'
'Ð'→'D'
'Ý'→'Y'
'Ÿ'→'Y'

Appendix3.

'DIPL.-CHEM. DR.RER.NAT.'

'DIPL.-CHEM. DR.-ING.'

'CHEMIE-ING. GRAD.'

'DR. DIPL. LANDWIRT'

'DIPL.-CHEM.,DR.'

'DIPL.-CHEM. DR.'

'DR.DIPL.-CHEM.'

'DR.-ING. MECH.'

'-ING. MECH.'

'DR.DIPL.-CHEM.'

'DIPL.-CHEM.'

'DIPL.-MATH.'

'DIPL.-PHYS.'

'DIPL.-ING.'

'ING.- GRAD'

'ING. GRAD.'

'DIPL.-BIO.'

'IR.-CHEM.'

'PROF. DR.'

'RER. NAT.'

'NAT.RER.'

'-INFORM.'

'DIPL-ING'

'LANDWIRT'

'DR.-ING.'

'PROF.DR.'

'RER.NAT'

'-CHEM.'

'DR.-MATH.'

'-MATH.'

'TECHN.'

'DR.-PHYS.'

'-PHYS.'	'PHIL.'
'DIPL.-'	'GRAD.'
'PH. D.'	'-BIO.'
'DIPL.'	'MED.'
'PROF.'	'-ING'
'PH.D.'	'ING.'
'-ING.'	'VET.'
'CHEM.'	'DR.'
'WIRT.'	'DR,'
'PHYS.'	'FH'

Appendix 4.

Austria (AT), Belgium (BE), Bulgaria (BG), Switzerland (CH), Cyprus (CY), Czech Republic (CZ), Germany (DE), Iceland (IS), Denmark (DK), Estonia (EE), Spain (ES), Finland (FI), France (FR), Greece (GR), Hungary (HU), Ireland (IE), Italy (IT), Lichtenstein (LI), Lithuania (LT), Luxemburg (LU), Latvia (LV), Malta (MT), the Netherlands (NL), Norway (NO), Poland (PL), Portugal (PT), Romania (RO), Sweden (SE), Slovenia (SI), Slovak Republic (SK), United Kingdom (UK).

Llista Document de Treball
List Working Paper

WP 2011/05 "Singling out individual inventors from patent data" Miguélez,E. and Gómez-Miguélez, I.

WP 2011/04 "¿La sobreeducación de los padres afecta al rendimiento académico de sus hijos?" Nieto, S; Ramos, R.

WP 2011/03 "The Transatlantic Productivity Gap: Is R&D the Main Culprit?" Ortega-Argilés, R.; Piva, M.; and Vivarelli, M.

WP 2011/02 "The Spatial Distribution of Human Capital: Can It Really Be Explained by Regional Differences in Market Access?" Karahasan, B.C. and López-Bazo, E

WP 2011/01 "I If you want me to stay, pay" . Claeys, P and Martire, F

WP 2010/16 "Infrastructure and nation building: The regulation and financing of network transportation infrastructures in Spain (1720-2010)"Bel,G

WP 2010/15 "Fiscal policy and economic stability: does PIGS stand for Procyclicality In Government Spending?" Maravalle, A ; Claeys, P.

WP 2010/14 "Economic and social convergence in Colombia" Royuela, V; Adolfo García, G.

WP 2010/13 "Symmetric or asymmetric gasoline prices? A meta-analysis approach" Perdiguero, J.

WP 2010/12 "Ownership, Incentives and Hospitals" Fageda,X and Fiz, E.

WP 2010/11 "Prediction of the economic cost of individual long-term care in the Spanish population" Bolancé, C; Alemany, R ; and Guillén M

WP 2010/10 "On the Dynamics of Exports and FDI: The Spanish Internationalization Process" Martínez-Martín J.

WP 2010/09 "Urban transport governance reform in Barcelona" Albalate, D ; Bel, G and Calzada, J.

WP 2010/08 "Cómo (no) adaptar una asignatura al EEES: Lecciones desde la experiencia comparada en España" Florido C. ; Jiménez JL. and Perdiguero J.

WP 2010/07 "Price rivalry in airline markets: A study of a successful strategy of a network carrier against a low-cost carrier" Fageda, X ; Jiménez J.L. ; Perdiguero , J.

WP 2010/06 "La reforma de la contratación en el mercado de trabajo: entre la flexibilidad y la seguridad" Royuela V. and Manuel Sanchis M.

WP 2010/05 "Discrete distributions when modeling the disability severity score of motor victims" Boucher, J and Santolino, M

WP 2010/04 "Does privatization spur regulation? Evidence from the regulatory reform of European airports . Bel, G. and Fageda, X."

WP 2010/03 "High-Speed Rail: Lessons for Policy Makers from Experiences Abroad". Albalate, D ; and Bel, G."

WP 2010/02 "Speed limit laws in America: Economics, politics and geography". Albalate, D ; and Bel, G."

WP 2010/01 "Research Networks and Inventors' Mobility as Drivers of Innovation: Evidence from Europe" Miguélez, E. ; Moreno, R. "

WP 2009/26 "Social Preferences and Transport Policy: The case of US speed limits" Albalate, D.

WP 2009/25 "Human Capital Spillovers Productivity and Regional Convergence in Spain" , Ramos, R ; Artis, M.; Suriñach, J.

WP 2009/24 "Human Capital and Regional Wage Gaps" ,López-Bazo,E. Motellón E.

WP 2009/23 "Is Private Production of Public Services Cheaper than Public Production? A meta-regression analysis of solid waste and water services" Bel, G.; Fageda, X.; Warner. M.E.

WP 2009/22 "Institutional Determinants of Military Spending" Bel, G., Elias-Moreno, F.

WP 2009/21 "Fiscal Regime Shifts in Portugal" Afonso, A., Claeys, P., Sousa, R.M.

WP 2009/20 "Health care utilization among immigrants and native-born populations in 11 European countries. Results from the Survey of Health, Ageing and Retirement in Europe" Solé-Auró, A., Guillén, M., Crimmins, E.M.

WP 2009/19 "La efectividad de las políticas activas de mercado de trabajo para luchar contra el paro. La experiencia de Cataluña" Ramos, R., Suriñach, J., Artís, M.

WP 2009/18 "Is the Wage Curve Formal or Informal? Evidence for Colombia" Ramos, R., Duque, J.C., Suriñach, J.

WP 2009/17 "General Equilibrium Long-Run Determinants for Spanish FDI: A Spatial Panel Data Approach" Martínez-Martín, J.

WP 2009/16 "Scientists on the move: tracing scientists' mobility and its spatial distribution" Miguélez, E.; Moreno, R.; Suriñach, J.

WP 2009/15 "The First Privatization Policy in a Democracy: Selling State-Owned Enterprises in 1948-1950 Puerto Rico" Bel, G.

WP 2009/14 "Appropriate IPRs, Human Capital Composition and Economic Growth" Manca, F.

WP 2009/13 "Human Capital Composition and Economic Growth at a Regional Level" Manca, F.

WP 2009/12 "Technology Catching-up and the Role of Institutions" Manca, F.

WP 2009/11 "A missing spatial link in institutional quality" Claeys, P.; Manca, F.

WP 2009/10 "Tourism and Exports as a means of Growth" Cortés-Jiménez, I.; Pulina, M.; Riera i Prunera, C.; Artís, M.

WP 2009/09 "Evidence on the role of ownership structure on firms' innovative performance" Ortega-Argilés, R.; Moreno, R.

WP 2009/08 "¿Por qué se privatizan servicios en los municipios (pequeños)? Evidencia empírica sobre residuos sólidos y agua" Bel, G.; Fageda, X.; Mur, M.

WP 2009/07 "Empirical analysis of solid management waste costs: Some evidence from Galicia, Spain" Bel, G.; Fageda, X.

WP 2009/06 "Intercontinental fligths from European Airports: Towards hub concentration or not?" Bel, G.; Fageda, X.

WP 2009/05 "Factors explaining urban transport systems in large European cities: A cross-sectional approach" Albalate, D.; Bel, G.

WP 2009/04 "Regional economic growth and human capital: the role of overeducation" Ramos, R.; Suriñach, J.; Artís, M.

WP 2009/03 "Regional heterogeneity in wage distributions. Evidence from Spain" Motellón, E.; López-Bazo, E.; El-Attar, M.

WP 2009/02 "Modelling the disability severity score in motor insurance claims: an application to the Spanish case" Santolino, M.; Boucher, J.P.

WP 2009/01 "Quality in work and aggregate productivity" Royuela, V.; Suriñach, J.

WP 2008/16 "Intermunicipal cooperation and privatization of solid waste services among small municipalities in Spain" Bel, G.; Mur, M.

WP 2008/15 "Similar problems, different solutions: Comparing refuse collection in the Netherlands and Spain" Bel, G.; Dijkgraaf, E.; Fageda, X.; Gradus, R.

WP 2008/14 "Determinants of the decision to appeal against motor bodily injury settlements awarded by Spanish trial courts" Santolino, M

WP 2008/13 "Does social capital reinforce technological inputs in the creation of knowledge? Evidence from the Spanish regions" Miguélez, E.; Moreno, R.; Artís, M.

WP 2008/12 "Testing the FTPL across government tiers" Claeys, P.; Ramos, R.; Suriñach, J.

WP 2008/11 "Internet Banking in Europe: a comparative analysis" Arnaboldi, F.; Claeys, P.

WP 2008/10 "Fiscal policy and interest rates: the role of financial and economic integration" Claeys, P.; Moreno, R.; Suriñach, J.

WP 2008/09 "Health of Immigrants in European countries" Solé-Auró, A.; M.Crimmins, E.

WP 2008/08 "The Role of Firm Size in Training Provision Decisions: evidence from Spain" Castany, L.

WP 2008/07 "Forecasting the maximum compensation offer in the automobile BI claims negotiation process" Ayuso, M.; Santolino, M.

WP 2008/06 "Prediction of individual automobile RBNS claim reserves in the context of Solvency II" Ayuso, M.; Santolino, M.

WP 2008/05 "Panel Data Stochastic Convergence Analysis of the Mexican Regions" Carrion-i-Silvestre, J.L.; German-Soto, V.

WP 2008/04 "Local privatization, intermunicipal cooperation, transaction costs and political interests: Evidence from Spain" Bel, G.; Fageda, X.

WP 2008/03 "Choosing hybrid organizations for local services delivery: An empirical analysis of partial privatization" Bel, G.; Fageda, X.

WP 2008/02 "Motorways, tolls and road safety. Evidence from European Panel Data" Albalate, D.; Bel, G.

WP 2008/01 "Shaping urban traffic patterns through congestion charging: What factors drive success or failure?" Albalate, D.; Bel, G.

WP 2007/19 "La distribución regional de la temporalidad en España. Análisis de sus determinantes" Motellón, E.

WP 2007/18 "Regional returns to physical capital: are they conditioned by educational attainment?" López-Bazo, E.; Moreno, R.

WP 2007/17 "Does human capital stimulate investment in physical capital? evidence from a cost system framework" López-Bazo, E.; Moreno, R.

WP 2007/16 "Do innovation and human capital explain the productivity gap between small and large firms?" Castany, L.; López-Bazo, E.; Moreno, R.

WP 2007/15 "Estimating the effects of fiscal policy under the budget constraint" Claeys, P.

WP 2007/14 "Fiscal sustainability across government tiers: an assessment of soft budget constraints" Claeys, P.; Ramos, R.; Suriñach, J.

WP 2007/13 "The institutional vs. the academic definition of the quality of work life. What is the focus of the European Commission?" Royuela, V.; López-Tamayo, J.; Suriñach, J.

WP 2007/12 "Cambios en la distribución salarial en españa, 1995-2002. Efectos a través del tipo de contrato" Motellón, E.; López-Bazo, E.; El-Attar, M.

WP 2007/11 "EU-15 sovereign governments' cost of borrowing after seven years of monetary union" Gómez-Puig, M..

WP 2007/10 "Another Look at the Null of Stationary Real Exchange Rates: Panel Data with Structural Breaks and Cross-section Dependence" Syed A. Basher; Carrion-i-Silvestre, J.L.

WP 2007/09 "Multicointegration, polynomial cointegration and I(2) cointegration with structural breaks. An application to the sustainability of the US external deficit" Berenguer-Rico, V.; Carrion-i-Silvestre, J.L.

WP 2007/08 "Has concentration evolved similarly in manufacturing and services? A sensitivity analysis" Ruiz-Valenzuela, J.; Moreno-Serrano, R.; Vaya-Valcarce, E.

WP 2007/07 "Defining housing market areas using commuting and migration algorithms. Catalonia (Spain) as an applied case study" Royuela, C.; Vargas, M.

WP 2007/06 "Regulating Concessions of Toll Motorways, An Empirical Study on Fixed vs. Variable Term Contracts" Albalate, D.; Bel, G.

WP 2007/05 "Decomposing differences in total factor productivity across firm size" Castany, L.; Lopez-Bazo, E.; Moreno, R.

WP 2007/04 "Privatization and Regulation of Toll Motorways in Europe" Albalate, D.; Bel, G.; Fageda, X.

WP 2007/03 "Is the influence of quality of life on urban growth non-stationary in space? A case study of Barcelona" Royuela, V.; Moreno, R.; Vayá, E.

WP 2007/02 "Sustainability of EU fiscal policies. A panel test" Claeys, P.

WP 2007/01 "Research networks and scientific production in Economics: The recent spanish experience" Duque, J.C.; Ramos, R.; Royuela, V.

WP 2006/10 "Term structure of interest rate. European financial integration" Fontanals-Albiol, H.; Ruiz-Dotras, E.; Bolancé-Losilla, C.

WP 2006/09 "Patrones de publicación internacional (ssci) de los autores afiliados a universidades españolas, en el ámbito económico-empresarial (1994-2004)" Suriñach, J.; Duque, J.C.; Royuela, V.

WP 2006/08 "Supervised regionalization methods: A survey" Duque, J.C.; Ramos, R.; Suriñach, J.

WP 2006/07 "Against the mainstream: nazi privatization in 1930s germany" Bel, G.

WP 2006/06 "Economía Urbana y Calidad de Vida. Una revisión del estado del conocimiento en España" Royuela, V.; Lambiri, D.; Biagi, B.

WP 2006/05 "Calculation of the variance in surveys of the economic climate" Alcañiz, M.; Costa, A.; Guillén, M.; Luna, C.; Rovira, C.

WP 2006/04 "Time-varying effects when analysing customer lifetime duration: application to the insurance market" Guillen, M.; Nielsen, J.P.; Scheike, T.; Perez-Marin, A.M.

WP 2006/03 "Lowering blood alcohol content levels to save lives the european experience" Albalate, D.

WP 2006/02 "An analysis of the determinants in economics and business publications by spanish universities between 1994 and 2004" Ramos, R.; Royuela, V.; Suriñach, J.

WP 2006/01 "Job losses, outsourcing and relocation: empirical evidence using microdata" Artís, M.; Ramos, R.; Suriñach, J.

IREA
Institut de Recerca en Economia Aplicada Regional i Pública
Research Institute of Applied Economics
Universitat de Barcelona
Av. Diagonal, 690 • 08034 Barcelona

WEBSITE: www.ub.edu/irea/ • **CONTACT:** irea@ub.edu