DOCUMENT DE TREBALL

XREAP2011-03

Singling out individual inventors from patent data

Ernest Miguélez (AQR-IREA)
Ismael Gómez-Miguélez

XARXA
DE REFERÈNCIA
EN ECONOMIA APLICADA

Singling out individual inventors from patent data[1]

Ernest Miguélez[†] & Ismael Gómez-Miguélez[§]

[†]AQR-IREA. Department of Econometrics, Statistics and Spanish Economy. University of Barcelona, Av. Diagonal 690, 08034 Barcelona, Spain. emiguelez@ub.edu.
[§] Signal Theory and Communications Department. Technical University of Catalonia, c/ Jordi Girona 1-3, 08034 Barcelona, Spain. ismael.gomez@tsc.upc.edu

Abstract

An increasing number of studies have sprung up in recent years seeking to identify individual inventors from patent data. Different heuristics have been suggested to use their names and other information disclosed in patent documents in order to find out "who is who" in patents. This paper contributes to this literature by setting forth a methodology to identify them using patents applied to the European Patent Office (EPO hereafter). As in the large part of this literature, we basically follow a three-steps 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 possible similar names; (3) the filtering stage, where additional information and different scoring schemes are used to filter out these potential same inventors. The paper includes some figures resulting of applying the algorithms to the set of European inventors applying to the EPO for a large period of time.

Key words: "Names game", patent data, unique inventors, name matching algorithms

JEL: C8, J61, O31, O33, R0

---

1. Introduction

Patent data offer a wide range of awesome information for research purposes in innovation economics, as well as regional economics and economic geography, among other social sciences' fields. In a patent document is contained information about the inventors' name of a patent, the owner's[2] name of the patent, the year and exact date of application of the patent[3], the exact addresses of both the inventors and the applicants, or the technological class to which the patent belongs. Further, by merging these datasets with patent citations, non-patent citation literature, or firm data, the information available is even larger and has helped us to better understand the ways in which knowledge is produced, exploited, diffused, and the like.

In spite of that patent data present serious caveats since not all inventions are patented, they do not have the same economic impact, and not all the patented inventions are commercially exploitable innovations (Griliches, 1991), they have been shown useful to proxy the inventiveness activity due to the fact that they do present the minimal standards of novelty, originality and potential profits, and must be a good proxy for economically profitable ideas (Bottazzi and Peri, 2003).

In such a setting, patent data have been widely used to analyse the innovation determinants of firms (Griliches, 1979; Hausman et al, 1984) or countries and regions, as well as to study the localized knowledge spillovers hypotheses, jointly with patent citation data (Bottazzi and Peri, 2003; Jaffe, 1986, 1989; Jaffe et al., 1993; Thompson and Fox-Kean, 2005). Furthermore, growth regressions have used patent data as a proxy for knowledge stocks or technological capability, especially since the advent of the endogenous growth theory (Romer, 1986, 1990; Anghion and Howitt, 1995). More recently, patent data have served as relational data through co-patenting information and the use, among other thinks, of social network analysis techniques (Bolconi et al., 2004; Fleming et al., 2007; Singh, 2005).

Among this huge amount of literature, something is partially missing. Thus, what has been less studied so far is the inventor herself: her personal characteristics, her linkages with other inventors or firms, and her labour and geographical mobility; as well as the

---

[2] The owner of a patent is the firm, institution, or individual who appears as the owner in the patent document –under the head "applicant". We will call it indistinctively in the present paper *owner*, *applicant*, or *assignee*.

[3] The priority year is the first year a patent was applied for worldwide.

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 because of the fact that patent data do not provide a consistent list of unique personal identifiers. Thus, unique IDs for each inventor and for anyone else are missing. Indeed, the information which is close to a sort of inventor's ID is her own name (name, middle name, surname, and so on). Therefore, names have been used to identify unique inventors. Needless to say that this procedure is also problematic for two main reasons. First, names and surnames contained in the patent document might well be spelled differently in each patent. Second, it is also possible that two patents, with exactly the same name (say, John Smith) do not belong to the same inventor.

To deal with these and related drawbacks, a large body of literature has sprung up in recent years (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., 2008; Raffo and Lhuillery, 2009; Trajtenberg et al., 2006; Thoma and Torrisi, 2007).[4] These authors have tried to contribute to the correct identification of unique inventors using basically their names, several 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, any methodology has shown its superiority to the others. Indeed, most of them have new advantages compared to the others, though a number of shortcomings as well. Our suggestions in the present inquiry strongly feed from this former literature, and try to contribute to enrich it at the same time. Thus, our aim here will be to pick up what, in our opinion, constitute the main advantages of these studies, while leaving aside their main shortcomings. The methodology developed will be applied to, first, a small sample of inventors which we will use as benchmark to test the goodness-of-fit of the approach, and second, to a large dataset of European patents applied by European inventors for a large period of time.

It is worthwhile to mention that some of the alluded researchers have recently joined efforts within the "Academic Patenting in Europe (APE-INV)" project led by KITES-Bocconi University. This project aims to put together a number of best practices to identify

---

[4] A brief summary about the different methodologies applied in these alluded studies, as well as the scope of their empirical application, is included in the appendix.

inventors from patent data. A summary of this project can be found is Lissoni et al. (2010)[5], where an updated survey of related studies can also be found.

In the next section, an explanation of the problematic faced and the solutions adopted will be described in detail. Broadly speaking, the aforementioned literature divides the procedure to identify inventors in three main stages (see Raffo and Luhllery, 2009). The first one deals with data cleaning, homogenisation and standardisation. The second stage matches the name of the inventors to form groups of patents potentially belonging to the same inventor. Exact or approximate name matching algorithms have been used indistinctly. Finally, within each group of patents, different heuristics and algorithms have been used to do pair-wise comparisons and assert if every pair of patents belongs 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 –which have been manually checked by Carayol and Cassi (2009). Section 4 shows the results of applying the methodology to the whole list of patents applied 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.

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

Patent data contain a huge amount of information very useful to undertake different analysis. Patent data do not, however, provide a consistent list of unique inventors' personal identifiers. In such a setting, it is necessary to turn into the inventor's name and surname reported in the patent itself. Unfortunately, two main problems arise in dealing with this strategy. The first occurs when the name (or surname) of the same inventor is spelled differently in 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 drawback, the literature suggests performing a list of algorithms aimed to identify unique inventors using their names and surnames, and other

---

[5] The following website contains all the information related to the APE-INV project: http://www.esf-ape-inv.eu/.

useful information, disclosed in the patent document. Following Raffo and Luhllery (2009), we divide the methodology to get the final data in three steps: parsing, matching, and filtering stages –according to the authors' terminology (Ibid.).

The parsing stage

What we need to do first is to clean up the fields of the correspondent database containing the name and surname of the inventor, as well as the field with their addresses. Equally, we would like to homogenise and standardise as much as possible the structure of each field and its content, in order to allow for comparisons between inventors.

For the case of the "inventors' name" field, we basically have proceeded in two main ways. First, we have corrected all the corrupted characters benefiting from the available work by Raffo and Luhillery (2009), from the CEMI's PATSTAT[6] Knowledge Base, "Ecole Polytechnique Fédérale de Lausanne" (http://wiki.epfl.ch/patstat/cleaning), as well as from Lars Tönqvist's typography (http://www.thesauruslex.com/typo/eng/enghtml.htm) concerning the encoding in HTML of foreign characters. The idea was to replace these types of characters by the corresponding characters from the Latin alphabet and easily legible by the name matching algorithm. Thus, for instance, the following changes have been made:

— '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)

And for the case of non-HTML-legible foreign characters, like vowels with different accents, swung dashes, dieresis, and so forth, they have been also modified. Few examples are:

— '&Aacute' is 'Á' and turns into 'A'
— '&Oslash' is 'Ø' and turns into 'O'

---

[6] PATSTAT stands for Worldwide Patent Statistical Database.

— '&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 have also changed all the non-corrupted accentuated characters for their non-accentuated counterparts, and the last cleaning-up task was to upper case all the characters; and drop slashes, hyphens, accents, dieresis, and the like. The whole list of changes made is presented in Appendix 2.

Secondly, we harmonise as much as possible the field. We did so by placing in different fields the surname(s) of the inventor, the first name, and the middle name. The idea was to use both the surname and the first name as the basis for the subsequent algorithm (see next subsection).

The middle name may include: the real middle name, or middle names, or the initials of them, or other kind of information like the inventors' affiliation, a surname modifier and so on. In fact, when surnames modifiers or the inventor's affiliation are present, we place them in separate fields and we use them as additional information to test whether or not a pair of records belongs to the same inventor. Concretely, we have placed in a separate field all the information contained in the inventors' name field preceded by 'C/O' as the potential affiliation of the inventor.[7] Moreover, we have extracted an arbitrary list of surnames' modifiers from this same field and we have placed them in a separate field as well -some of them are 'Prof.', 'Dr.', 'Prof.-Dr.', 'Ing.', 'Jr.', 'PhD.', 'Chem.', and the whole list of surnames' modifiers is found in Appendix 3.

Concerning inventors' addresses, the cleaning-up process resembles the inventor's name counterpart – regarding corrupted characters and so on. With regards to harmonisation of fields, we proceed by placing in different field the single address (name of the street and building number), the zip code, and the name of the city. These three fields are going to be used in the filtering stage.

---

[7] 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'.

Moreover, additional information is retrieved from REGPAT. As pointed out elsewhere (Lissoni et al., 2010), any but one of the papers reviewed in the appendix section (Lai et al., 2009) makes use of information non-reported in PATSTAT or USPTO files. Thus, aside from the raw data extracted from REGPAT –which is shared with PATSTAT- we make use of the work made by the OECD within this alluded database. Even though PATSTAT users usually have access to country codes linked to inventors' and applicants' patents, supplementary information regarding a more refined spatial level from where the patent comes from is left to the researcher's search. Contrarily, additional information can be found in REGPAT. As explained in Maraut et al. (2008), they have used the address field of both inventors and applicants of patents to link them to micro-regions in OECD countries. For the case of Europe –which is our concern in the present research project- patents have been assigned to NUTS3[8] regions. Basically, the zip codes contained in that field are isolated and used to link them to the latest version of NUTS classification code (which corresponds to 2006). When the zip code is missing in the field, city's name is used instead. From the NUTS3 codes, one can easily retrieve the NUTS2 code to use them in the final stage of the present methodology.

The name matching stage

As said earlier, most of the algorithms found in the literature are based upon the inventors' name and surname to decide "who is who" in the "names game". However, even after cleaning, standardising, and harmonising these fields, it is possible to find two inventors' name string truly belonging to the same guy that are assigned to different people because different spelling –because of errors, for instance. Thus, the second step consists on codifying the strings of the mentioned fields in order to minimize these spelling problems which have introduced variations of the same inventor name –the name matching algorithm will help us, therefore, to minimize the **Type I error**[9].

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 might refer to character variations, including capitalisation (Trippl *versus* trippl), punctuation (López Bazo *versus* López-Bazo), spacing

---

[8] NUTS stands for the French acronym *"Nomenclature des Unités Territoriales Statistiques"*.
[9] 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.

(ERNESTMIGUELEZ *versus* ERNEST MIGUELEZ), or qualifiers (Rosina Moreno *versus* Prof. Dr. Rosina Moreno). Some of these mentioned sources of problems might be solved through the previous stage. However, other sources of mistakes might refer to spelling variations, including *insertion* (McCann *versus* MacCann), *omission* (Iammarino *versus* Iamarino), *substitution* (Maier *versus* Mayer), or *transposition* (Fingelton *versus* Fingleton). And finally it might refer to phonetic variations (Cooper in English would be spelled Cuper in German).

A name matching system must deal not only with spelling and phonetic concerns, but also with cultural aspects (Snae, 2007). For instance, there exist spelling analysis-based algorithms (like the Guth and Levenshtein alogarithms), 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 recognise, 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 used to list all the individuals in the US census records starting from 1880. It encodes 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 has finished before using the whole number of digits. 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 surname Soundex-code and the name Soundex-code we build what Trajtenberg et al. (2006) called the *p-sets* (potentially the same inventor). Each different *p-set* is therefore identified as a different, unique inventor. In this way, we encode, with the same Soundex-code, the strings that differ slightly but actually belong to the same person (like those of the former examples). Notwithstanding, this procedure might 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 belong 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 –this is done in order to minimize **Type II errors**[10]. The approach chosen in this stage is close to Lissoni et al.'s (2006) methodology, as well as the work by Trajtenberg et al. (2006).

We have run as much tests as the raw data permit, squeezing all the information linked to each patent in order to optimise the identification procedure. We then assign an arbitrary score to each comparison made, and we add up total scores for every pair-wise comparison. This results in the "similarity score" for pairs of inventors with the same Soundex code. Afterwards, we compare it with a pre-determined numerical threshold – up to which we decide if two records belong to the same inventor or not. After doing 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.

---

[10] The **"Type II errors"** are those incurred when we end up matching records that belong in fact to different inventors.

The code to run the pair-wise comparisons was written with Java using the Netbeans software.[11] In Table 2 of 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 said before, patent documents information is stored in different databases, being PATSTAT the original one. Conversely, we have used the information stored in REGPAT database, prepared by the OECD. REGPAT contains basically the same information as PATSTAT. It includes, however, information from the region to which the inventors' addresses reported in the document corresponds. The NUTS3 code is then included, from which we can easily retrieve the NUTS2 code, if necessary. As for the applicant is concerned, we have used data from the KITES-PatStat database (Bocconi University – Milan). What the KITES group have done with the applicants data is to give a code to each firm trying to avoid, on the one side, spelling problems as well as corrupted characters problems; and on the other side, giving the same code to different applicants' names when they were actually the same applicant. Thus, for instance, the same code was given to 'I.B.M.' and to 'International Business Machines'. Additionally, KITES gives a group code to each patent if it can be retrieved from 'Dun&Bradstreet'. The idea is that in few cases, different applicants might belong to the same corporative group, and therefore this information can be used to identify inventors.[12] Citation data to test if one inventor cites the other one is taken from the 'OECD EP/WO Citation database', which stores citation data also contained in patent documents. Below, the complete list of tests run is given:

- **Inventor's bibliographical information**
  - *Same middle name (encoded using Soundex with 6 digits)*
  - *Same inventors' 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*

- o *Same NUTS3 region code*
- o *Same NUTS2 region code*
- **Information from the patent itself: applicant(s) and technological class(es)**
  - o *Same applicant code (according to the KITES-PatStat codification)*
  - o *Same company code (according to the KITES-PatStat codification)*
  - o *Same group code (according to the KITES-PatStat codification)*
  - o *Same technological class(es) –IPC code (4 digits)*
  - o *Same technological class(es) –IPC code (6 digits)*
  - o *Same technological class(es) –IPC code (12 digits)*
- **Citations information**
  - o *If one patent cites the other*

3. Testing the algorithms: The benchmark dataset

Once the three-step methodology is designed, one should go to implement it on real patent data. The main problem is that we are completely unable to ascertain whether or not the methodology suggested in the present study (as well as other similar methodologies shown elsewhere) is good enough to identify individual inventors. In trying to overcome these shortcomings, we are going to use a sample that has been checked manually. Using this benchmark, we are going to decide a scoring scheme that will give us the highest goodness-of-fit, and we will apply this same scoring scheme (and threshold) to the whole dataset. We acknowledge, however, that this procedure is dependent upon the "quality" of the benchmark, that is, to what extent this benchmark is truly representative of the whole dataset.

The benchmark used is that by Carayol and Cassi (2009), to which we have had access thanks to our participation in the APE-INV project. Obviously, we are indebt to them for their invaluable work based on checking, by hand, the sample.

The French academic inventors' benchmark

This benchmark is made up of 424 French academic inventors (see Lissoni et al., 2010; and Lissoni et al., 2008; for an in-depth description), affiliated to French universities during 2004-2005. This set of inventors is the result of matching EPO patents from 1975 to 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. By-hand checking of the total number of patents belonging to each one of these academics has been also performed 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.

Goodness-of-fit measures and undertaken approach

Before going further, we show below 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:

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

Where:
- **True Positives** are each couplet of patents belonging to a given same inventor in the benchmark that are said to belong as well to the same inventor as the result of the algorithm.
- **False Positives** are each couplet of patents not belonging to a given same inventor in the benchmark that are said to belong to the same inventor as the result of the algorithm.
- **False Negatives** are each couplet of patents belonging to a given same inventor in the benchmark that are not said to belong to the same inventor as the result of the algorithm.
- And, for information, **True Negatives** are each couplet of patents not belonging to a given same inventor in the benchmark that are said to not belong either to the same inventor as the result of the algorithm.

We turn now to the description of our approach. As well known, one of the main problems in this type of exercise is the decision about the weights we should assign to each of the characteristics tested. Former studies do not have a common pave to follow, and some of them give a more or less homogeneous score to each test (Lissoni et al., 2006). Others give different scores according to an (arbitrary) importance given to each test (Trajtenberg et al., 2006), whilst some other examples limits their methodology to 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 trying to keep things simply, what we are going to do here is to start with a homogenous scoring scheme –as in Lissoni et al. (2006). Afterwards, we will give different values to one of the parameters, concretely the threshold up to which a given pair of records is said to belong to the same inventor, and we will present the results for 25 different thresholds. We have repeated this same procedure using different scoring schemes, by giving heterogeneous scores to the tests, according to previous studies (Agrawal et al., 2006; Trajtenberg et al., 2006), as well as our own common sense. None of these alternative scoring schemes can be said to be superior to the former one -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 benchmark, using the scoring scheme detailed in Table 2 and different thresholds. As can be seen, the precision and recall rates are very high. They also allow us to choose the threshold that better suits our purposes, given a scoring scheme. In Figure 1 below are depicted points resulting from the combination of recall and precision rates.

Figure 1. Goodness-of-fit: recall and precision rates
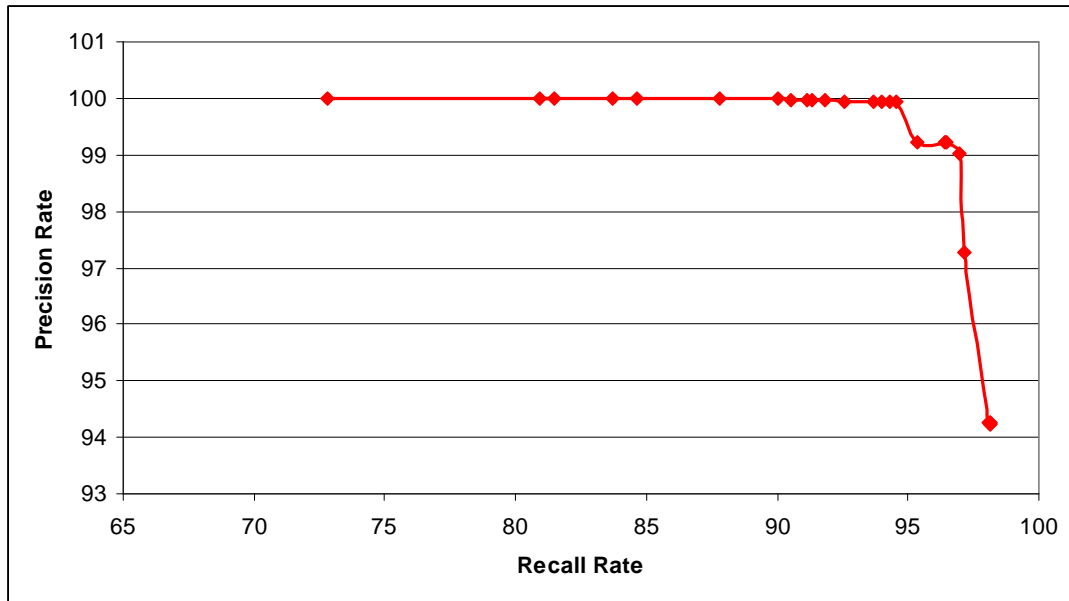
Given that the main purpose of subsequent econometric estimations is the study of labour and geographical mobility of inventors, we are especially interested in minimizing the number of false positives –each couplet of patents not belonging to the same inventor in the benchmark that are said to belong to the same inventor as the result of the algorithm- but without compromising the number of false negative. Consequently, given the aforementioned scoring scheme, by setting the threshold at 15 we have a very limited number of false positives (4) and the lowest number of false negatives among the thresholds with only 4 false positives.

Table 3. Results French benchmark for different thresholds

True Positives	True Negatives	False Positives	False Negatives	Threshold	Precision Rate	Recall Rate
17062	3963590	1042	326	0	98.13	94.24
17062	3963590	1042	326	1	98.13	94.24
17062	3963592	1040	326	2	98.13	94.25
17062	3963592	1040	326	3	98.13	94.25
17056	3963592	1040	332	4	98.09	94.25
16888	3964160	472	500	5	97.12	97.28
16856	3964464	168	532	6	96.94	99.01
16776	3964502	130	612	7	96.48	99.23
16764	3964502	130	624	8	96.41	99.23
16578	3964502	130	810	9	95.34	99.22
16438	3964620	12	950	10	94.54	99.93
16392	3964620	12	996	11	94.27	99.93
16344	3964622	10	1044	12	94.00	99.94
16292	3964622	10	1096	13	93.70	99.94
16094	3964622	10	1294	14	92.56	99.94
15970	3964628	4	1418	15	91.84	99.97
15878	3964628	4	1510	16	91.32	99.97
15850	3964628	4	1538	17	91.15	99.97
15740	3964628	4	1648	18	90.52	99.97
15658	3964632	0	1730	19	90.05	100.00
15270	3964632	0	2118	20	87.82	100.00
14720	3964632	0	2668	21	84.66	100.00
14558	3964632	0	2830	22	83.72	100.00
14072	3964632	0	3316	23	80.93	100.00
14172	3964632	0	3216	24	81.50	100.00
12662	3964632	0	4726	25	72.82	100.00

4. Whole patent dataset and descriptive statistics

In this section, we show the application of the methodology described so far to the whole dataset of patents. Concretely, we apply the procedures to the REGPAT database (OECD, January 2010 edition). We first briefly describe the data used, alongside a number of figures. We present then a summary of results in terms of inventors identified, their average characteristics, their technological and spatial distributions, and their temporal evolution.

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 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. (2006) for a methodological note. Thanks to their fruitful work, we can identify the region from which each inventor works when she applies for a patent. Basically, they are concerned with the process of

regionalisation of patent data at very low levels of disaggregation, which they assess using the addresses of the inventor documented in patent documents – the ZIP code or, in its absence, the town name. This regionalisation procedure provides researchers with a complete dataset of patents applied for under the European Patent Office, containing a rich amount 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 about 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 take account of co-authorships and multi-applicants, and finally the technological class(es) to which each patent corresponds.

Since our final aim is a regional aggregated approach, we have restricted our identification methodology to those inventors who live in European countries. The whole list of countries is shown in the Appendix 4. From a time dimension perspective, we have exploited all the data available and hence we have data from 1978 to 2005. According to Maraut et al. (2008), the regionalisation process undertaken by the OECD reached a success rate of 98% for the case of EPO patents. However, for some countries this processes ended up in allocations of NUTS codes with a breakdown –for the case of Germany, for instance, the share of addresses with a breakdown in different NUTS3 is around 14% (Ibid.). Since our prime interest is a correct regionalisation to study mobility across regions, we remove all the patents with a regionalisation breakdown below 70%. Additionally, for some addresses no allocation is reached, for various reasons: town names allocated in different NUTS3 regions, addresses referring to a wrong country, the address field is empty or not valid, and the like. We also remove all these patents. All in all, however, the number of records eliminated for these several reasons do not exceed the 1.8%. Our final dataset contains 2,297,196 records, which corresponds to all the pair-wise combinations of inventors' name strings plus patent number, from 1978 to 2005. This corresponds to 1,041,080 different patents, meaning an average number of different inventors per patent around 2.21. The distribution of EPO patents across countries is very unbalanced, as can be seen from Figure 2, being Germany the most productive country in terms of innovation outputs, followed by France and Great Britain, irrespective on how patents are aggregated –fractional counts or full counts. Conversely, Malta is in the tail of the distribution.

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

Patents.
Fractional counts and full counts.
1975-2008

☐ PATSFRAC
■ PATSCOUNTS
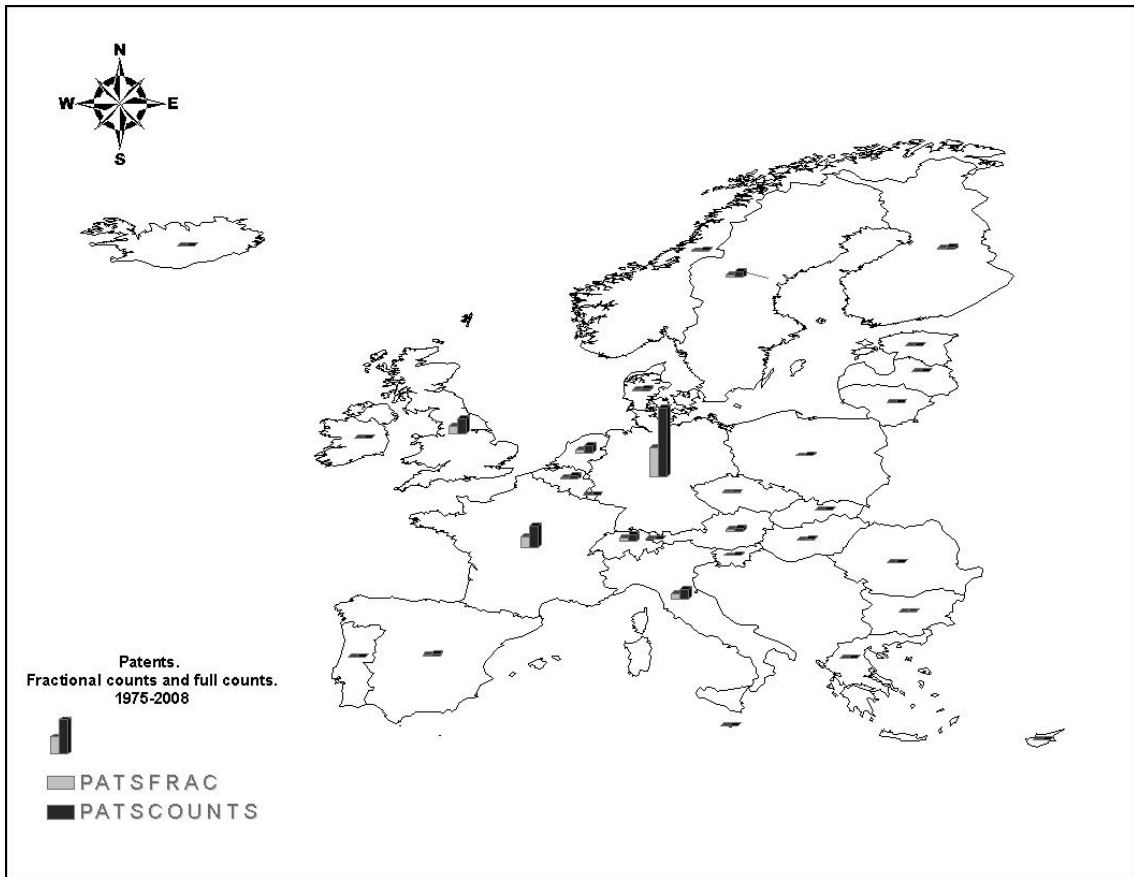
Additionally, this uneven distribution remains practically unchanged through time if we look at different time spans in different and separate time periods. This can be seen in Figure 3, where the distribution of patents across countries in different moments of time - 20 years gap- is depicted in maps.

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

i) 1981-1985   **ii) 2001-2005**
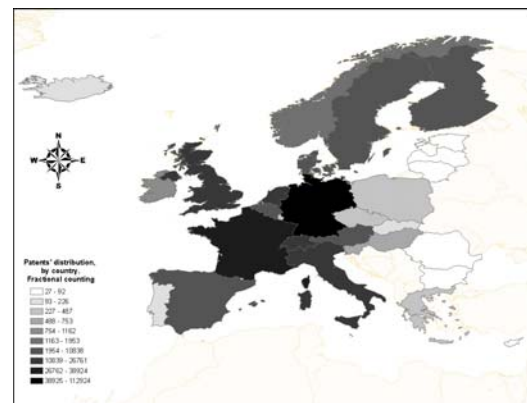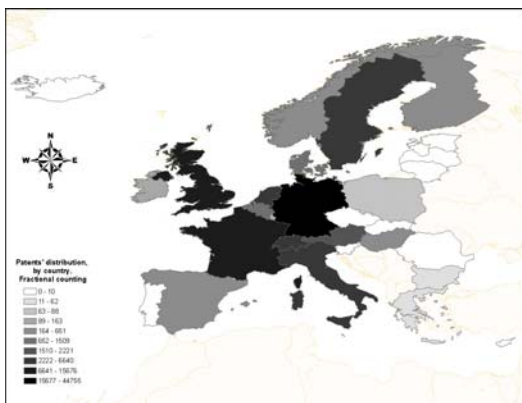
The evolution of patent activity over the mentioned sample shows a continuous increasing trend in the number of patent applications throughout the whole period. Few exceptions are the recession period experienced in the beginning of the nineties, and a small stagnation in the production of patents between 2001 and 2002 –coinciding with the "dot-com bubble". In any case, the overwhelming general increment in patent production may well be explained both by the rising technological complexity of economic activities, as well as the increasing use of the European Patent Office against to, or in complement with, national offices.

Figure 4. Patents' evolution in Europe, fractional counts. 1977-2005
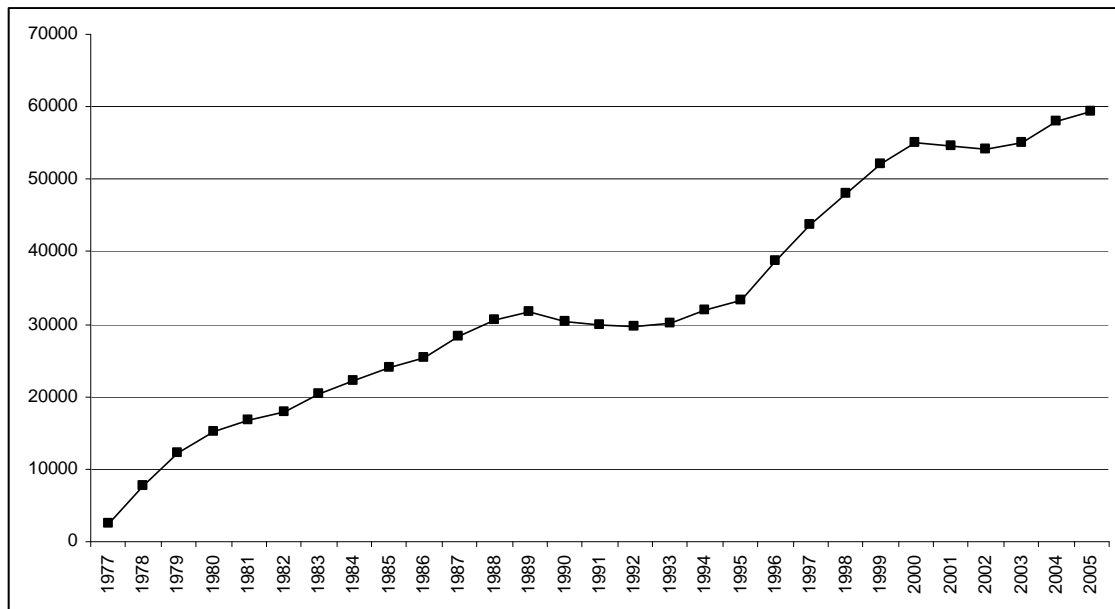
The distribution of patents across the space is even more unbalanced if we look at the regional level (at NUTS2 level of regional desegregation). Figure 5 depicts two maps corresponding to the regional distribution of patents in separate moments in time. As we can see, this distribution is very uneven as well, and in some of the cases it is even within countries –for the case of Great Britain or Spain, for instance. Regarding the time dimension, more regions show dark shades in the second period than in the first one, though 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
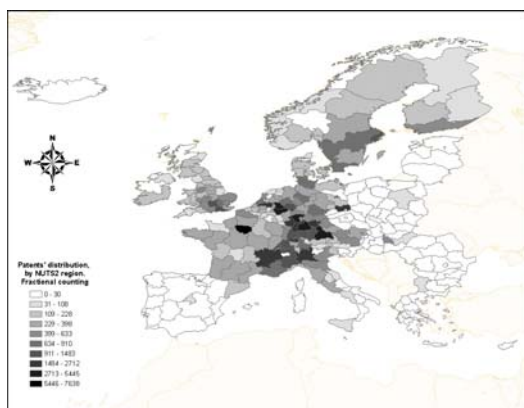
ii) 2001-2005
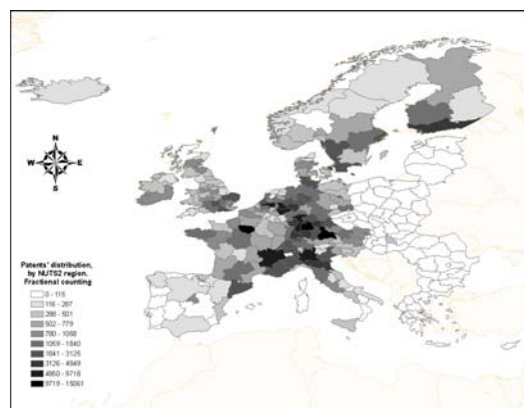
Results of the different stages of the methodology

The parsing stage

After the parsing stage -cleaning, harmonising and standardising the inventor's name field and the address field- few figures can be highlighted. Thus, the initial 2,297,196 records are made up of 29,017 different names, 257,227 surnames, and 678,324 combinations of name and surname. Additionally, 509,597 over 2,297,196 records (22.18%) have a middle name (or the initial of it). In 300,523 cases (13.08%) there exist a surname modifier, and in 30,262 records (1.32%), the affiliation of the inventor can be retrieved. In the following table, the most common names, surnames, and combination of both are presented.

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

Name	# records	Surname	# records	Name+Surname	# records
PETER	50058	MULLER	10758	EBERHARD AMMERMANN	526
JEAN	48213	SCHMIDT	7289	VOLKER REIFFENRATH	481
HANS	47832	FISCHER	5210	ROBERT SCHMIDT	473
MICHAEL	37625	SCHNEIDER	4761	HEINZ FOCKE	446
THOMAS	33710	WEBER	3825	HANS SANTEL	406
WOLFGANG	29232	MEYER	3586	GISELA LORENZ	381
KLAUS	28673	BAUER	3142	KLAUS MULLER	377
MARTIN	22362	WAGNER	3058	HANS MULLER	346
KARL	21218	MARTIN	2838	JEAN GUERET	344
ANDREAS	20753	SMITH	2792	SIEGFRIED STRATHMANN	340

As for the case of the addresses, it is worth to be highlighted that records are distributed in 127,131 different zip codes, 151,582 cities, 1,312 NUTS3 regions, and 289 NUTS2 regions.

In Table 5 below, the most repeated zip codes, cities, NUTS3 and NUTS2 in terms of number of record are reported.

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

Zip code	# records	City	# records	NUTS3	# records	NUTS2	# records
5656	40019	MUNCHEN	43597	NL414	49120	FR10	136638
8000	20003	EINDHOVEN	35531	FR101	38356	DE21	105090
8501	7478	PARIS	33611	DE212	35132	DE11	97669
1000	7456	BERLIN	26881	ITC45	30364	DE71	92653
5000	6605	STUTTGART	15004	FR105	28974	DEA1	85845
5090	6590	HAMBURG	13622	DE300	27107	DEA2	76701
5600	5630	KOELN	13362	SE110	24703	DEB3	67021
6700	5501	LEVERKUSEN	11537	CH040	23873	DE12	59475
4000	5157	MILANO	11446	DE115	22628	NL41	57010
75008	5139	DUSSELDORF	11334	FR103	20648	FR71	52932

The matching stage

After applying the name matching algorithm, that is, the Soundex code for names and surnames, several points must be highlighted. Recall from the former sections that this algorithm allow us to avoid spelling problems that introduced variation in the inventors' name field even if a given pair of records belongs to the same inventor. Unfortunately, however, this algorithm will force 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, might be due to completely different names, or due to misspellings of the same name. The same Table 6 below includes few examples of both cases for the case of the most frequent Soundex code. Again 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

All in all, as a result of applying the three stages using patent data from REGPAT OECD databases (January 2010 edition) we have finally identified 768,810 inventors from a sample of 2,297,196 initial records. This means an average of 2.99 patents per inventor, which is in line with similar studies in this field (see, for instance, Trajtenberg et al., 2006). As can be seen from Table 7 below, the distribution of the number of patents per inventor is very skewed, since the majority of inventors have only 1 patent (55.99% of them) or less than 6 patents (88.69%). In the meanwhile, 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 these identified inventors across countries is also very uneven. As expected, Germany is the leading country in hosting inventors (as it was the case for patents), followed by France and the UK, as can be seen from table 8 and figure 6.[13] On

---

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

the other side, Malta is the country hosting a lower number of inventors during the whole period under scrutiny.

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 distribution of inventors across the space is further confirmed in the following maps (Figure 6) where both the distribution of inventors over population is depicted both at country level (i) and at the NUTS2 level (ii).

Figure 6. Distribution of inventors across countries and NUTS2 regions.
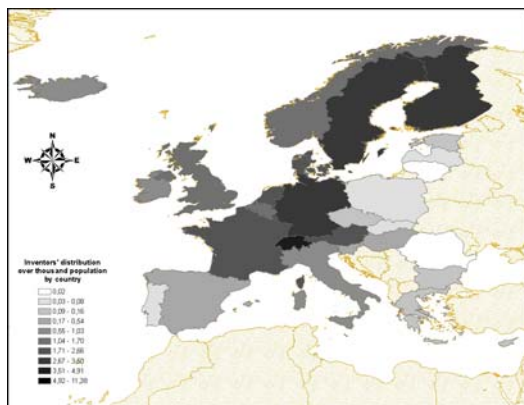
i) NUTS0   **ii) NUTS2**
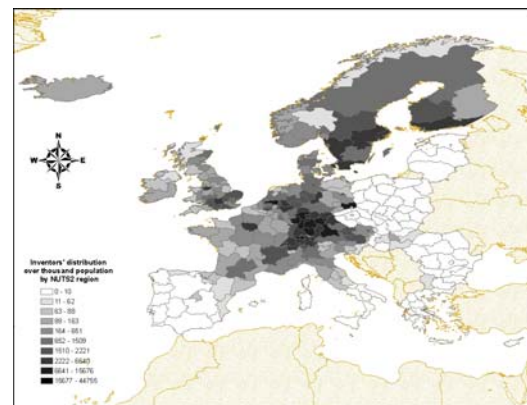
Note: To calculate this ratio, we have computed all the inventors identified throughout the whole period of analysis over population in 2005.

Figure 7 below shows the time evolution 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 as well as their time evolution is very

dependent upon the number of patents applied to the EPO. At the same time, however, spatial distribution and time evolution of patent applications are very 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 worthwhile itself.

Figure 7. Inventors' evolution in Europe. 1977-2005
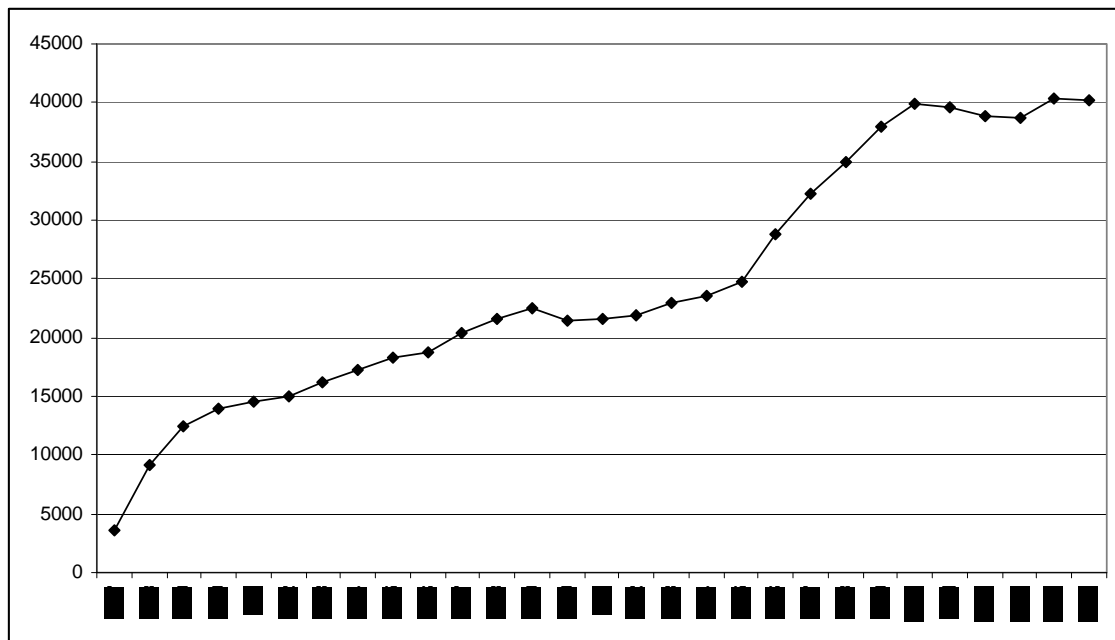
Another interesting point is related to the distribution of inventors across technological sectors.[14] Figure 8 below shows this distribution across technologies for the whole period under analysis (1977-2005). As can be seen, industrial processes, mechanical engineering, and electrical engineering are the sectors with more inventors. However, and contrary to their spatial distribution, the differences across technological sectors are not that pronounced.

---

[14] As regards the technological classification used to describe the distribution of inventors across technological sectors, we have adopted a technology-oriented classification, jointly elaborated 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.

Figure 8. Inventors' distribution across technological sectors. 1977-2005
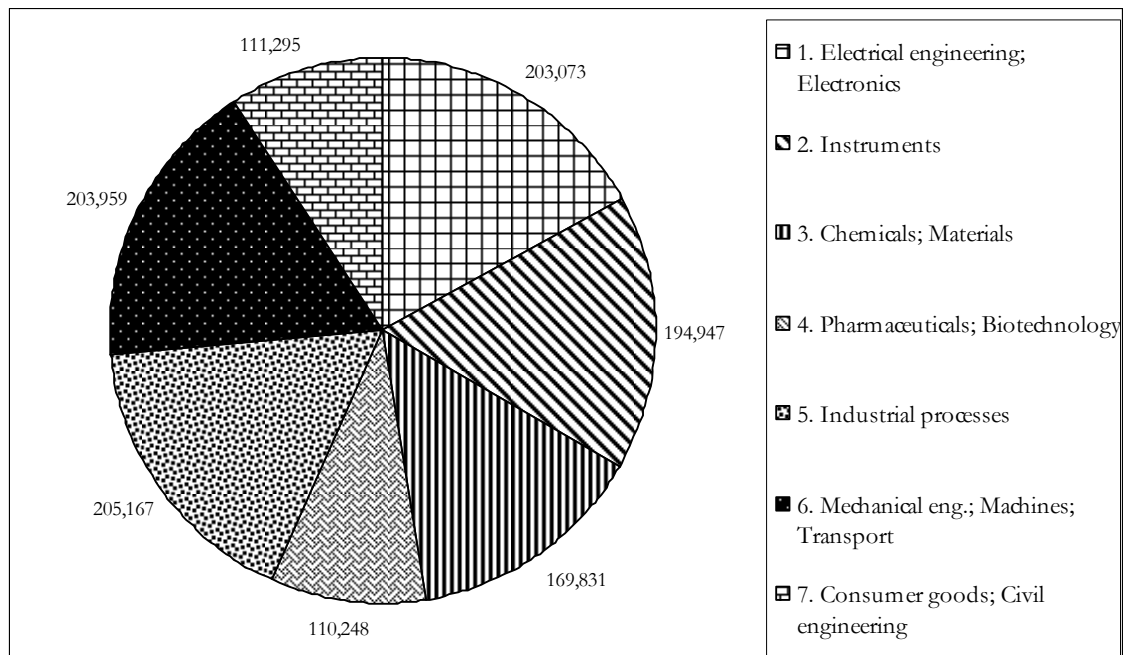
The following figures (Figure 9 and Figure 10) also show the evolution of inventors in time across different sectors. In spite of the growing tendency in the quantity of inventors in all 7 sectors, their relative importance has slightly changed during the whole period. Thus, although their respective share remains stable throughout time (Figure 10), several changes might be reported. Basically, one might observe that, through years, sectors like electrical engineering and pharmaceuticals and biotechnology have increased its importance, whilst industrial processes has decreased it. However, the number of inventors has sharply increased in all the sectors.

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

Legend:
- 1. Electrical engineering; Electronics
- 2. Instruments
- 3. Chemicals; Materials
- 4. Pharmaceuticals; Biotechnology
- 5. Industrial processes
- 6. Mechanical eng.; Machines; Transport
- 7. Consumer goods; Civil engineering
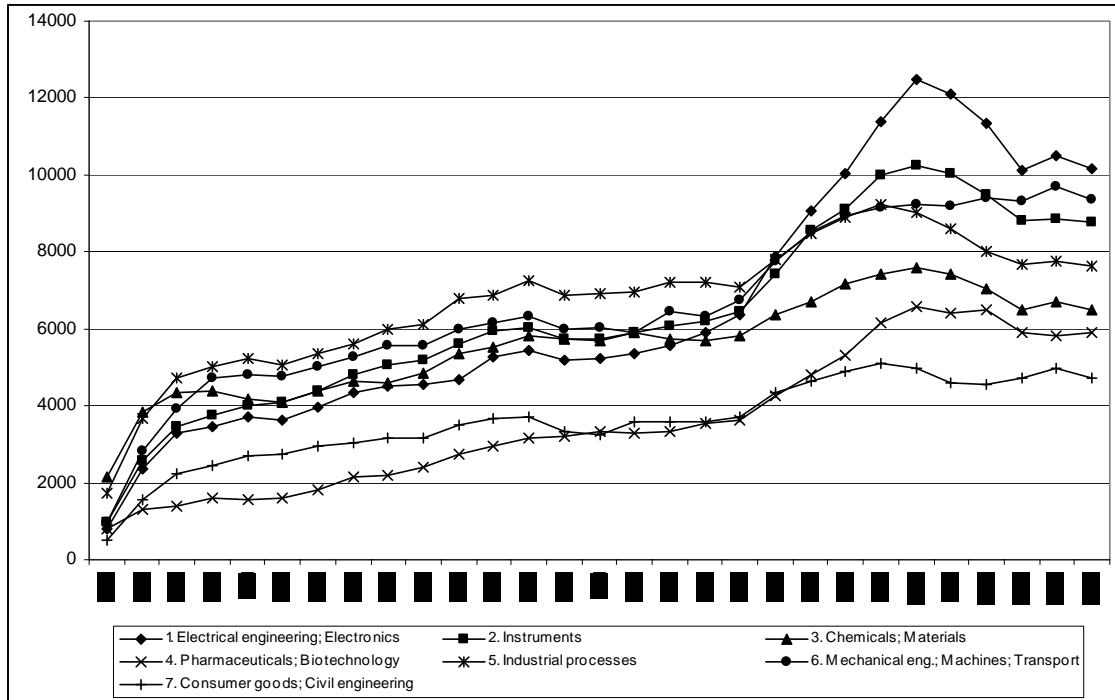
Figure 10. Inventors' distribution across technological sectors and time periods. 1977-2005

Legend:
- 7. Consumer goods; Civil engineering
- 6. Mechanical eng.; Machines; Transport
- 5. Industrial processes
- 4. Pharmaceuticals; Biotechnology
- 3. Chemicals; Materials
- 2. Instruments
- 1. Electrical engineering; Electronics
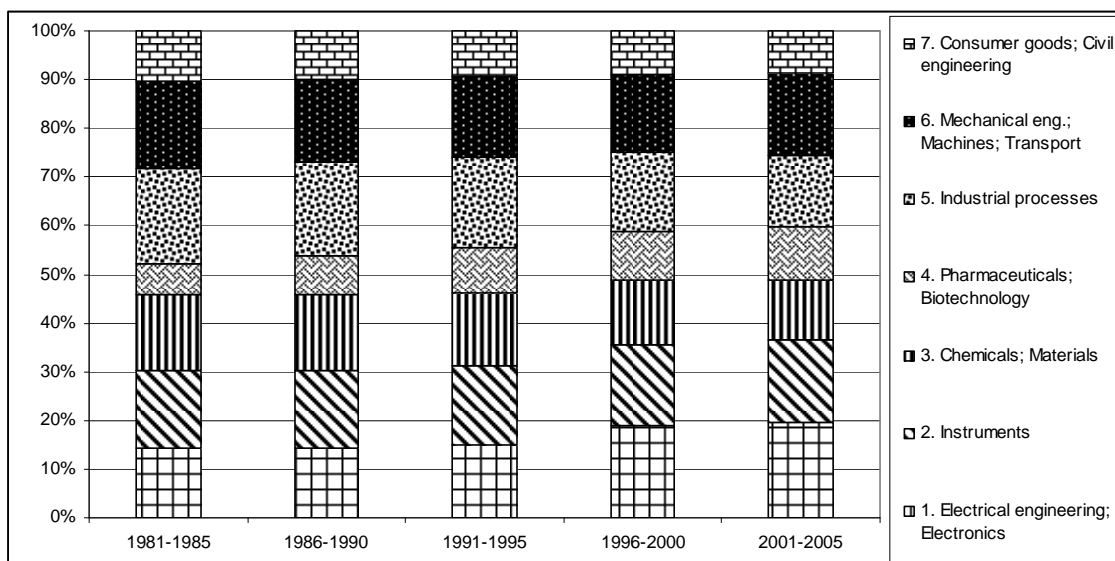
5. Conclusions

In the present paper we have described in a detailed fashion the methodology carried out to identify individual inventors using patent documents. To recap, this methodology consists on three steps. First, a cleaning-up process of the raw data; second, the use of SOUNDEX, a name matching algorithm to group possible similar names; and third, a

"splitting" algorithm to ascertain if every pair 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 assigned a score to each test and then we sum the scores up. If the total score reach a minimum threshold, a given couple of inventors were said to be the same person. In order to choose the scores we run iteratively our algorithm for a small sample of French academic inventors for whom we knew exactly "who is who". We have calculated recall and precision rates (false positives and false negatives) from this benchmark, and we have used the scoring scheme and threshold which best suits our purposes.

The way in which we have chosen the scores, however, is not free of criticism, due to the fact that we were not able to run all the possible combinations of scores and thresholds using all the tests performed. Thus, as a line of future research, we are planning to design an algorithm capable to decide endogenously the scores of the splitting algorithm by itself (this is done somehow by Carayol and Cassi, 2009).

References

Acs Z., Anselin L. and Varga A. (2002) Patents and innovation counts as measures of regional production of new knowledge, Research Policy 31, 1069–85;

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

Almeida P, Kogut B (1999) Localisation of knowledge and the mobility of engineers in regional networks. *Management Science* 45: 905-917

Anselin L, Varga A, Acs Z (1997) Local Geographic Spillovers between University Research and High Technology Innovations. *Journal of Urban Economics* 42: 422-448

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

Breschi S. and Lissoni F. (2009) Mobility of skilled workers and co-invention networks: an anatomy of localized knowledge flows, Journal of Economic Geography 9, 4, 439-68;

Crespi G, Geuna A, Nesta L (2007) The mobility of university inventors in Europe *Journal of Technology Transfer* 32(3): 195-215

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

Hoisl K (2009) Tracing mobile inventors: The causality between inventor mobility and inventor productivity. *Research Policy* 36(5): 615-636

Hoisl K (2007) Does mobility increase the productivity of inventors? *Journal of Technology Transfer* 34: 212-225

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

Jaffe A. B. (1989) Real effects of academic research, American Economic Review 79, 5, 957-70;

Jaffe AB, Trajtenberg M, Henderson R (1993) Geographic localisation of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics* 108: 577-598

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

Lenzi C (2009) Patterns and determinants of skilled workers' mobility: evidence from a survey of Italian inventors. *Economics of Innovation and New Technology* 18(2): 161-179

Lissoni F (2008) Academic inventors as brokers: An exploratory analysis of the KEINS database *CESPRI Working Paper*, 213

Lissoni F, Llerena P, McKelvey M, Sanditov B (2008) Academic patenting in Europe: new evidence from the KEINS database. *Research Evaluation* 16: 87–102

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

Miguelez. E.; R. Moreno and J. Suriñach (2009) "Inventors on the move: tracing inventors' mobility and its spatial distribution" IREA Working Papers 2009/16

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

Thoma, G., Torrisi, S., Gambardella, A., Guellec, D., Hall, B.H., Harhoff, D. (2009a) Methods and software for the harmonization and combination of datasets: A test based on IP-related data and accounting databases with a large panel of companies at the worldwide level, *mimeo*

Thoma, G. et al. (2009b) "Harmonizing and Combining Large Datasets: An Application to Patent and Finance Data". STI Working Paper, OECD, Paris.

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

Trajtenberg M, Shiff G (2008) Identification and mobility of Israeli patenting inventors, The Pinhas Sapir Center for Development, Tel Aviv University DP No. 5-2008

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 metropolitant 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, Sweeden 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'
'Â³'→'o'
'Â²'→'o'
'Ã´'→'o'
'Ã '→'OE'
'Ã–'→'OE'
'Â¶'→'oe'
'Ã·'→'oe'
'Ã"'→'O'
'Ã¸'→'o'
'Ã˜'→'O'
'Ã“'→'O'
'ÃŸ'→'ss'
'Â·'→'u'
'Ãº'→'u'
'Â»'→'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'

		Accents, slashes, diaeresis, and other punctuation symbols:
'ô'→'o'	'ł'→'l'	'Ä'→'A'
'Œ'→'OE'	'Ń'→'N'	'Ë'→'E'
'œ'→'oe'	'ń'→'n'	'Ï'→'I'
'Û'→'U'	'Ś'→'S'	'Ö'→'O'
'û'→'u'	'ś'→'s'	'Ü'→'U'
'Ÿ'→'Y'	'Ź'→'Z'	'À'→'A'
'Ź'→'y'	'ź'→'z'	'È'→'E'
'ß'→'B'	'Ż'→'Z'	'Ì'→'I'
'Ő'→'O'	'ż'→'z'	'Ò'→'O'
'ő'→'o'	'Ã'→'A'	'Ù'→'U'
'Ű'→'U'	'ã'→'a'	'Á'→'A'
'ű'→'u'	'ª'→'a'	'É'→'E'
'Þ'→'P'	'º'→'o'	'Í'→'I'
'þ'→'p'	'Ă'→'A'	'Ó'→'O'
'Ā'→'A'	'ă'→'a'	'Ú'→'U'
'ā'→'a'	'Ş'→'S'	'Â'→'A'
'Ē'→'E'	'ş'→'s'	'Ê'→'E'
'ē'→'e'	'Ţ'→'T'	'Î'→'I'
'Ģ'→'G'	'ţ'→'t'	'Ô'→'O'
'ģ'→'g'	'¡'→''	'Û'→'U'
'Ī'→'I'	'¿'→''	'Î'→'I'
'ī'→'i'	'€'→''	'{'→' '
'Ķ'→'K'	'&pound'→''	'}'→' '
'ķ'→'k'	'«'→''	'('→' '
'Ļ'→'L'	'»'→''	')'→' '
'ļ'→'l'	'•'→''	'Ç'→'C'
'Ņ'→'N'	'†'→''	'Å'→'A'
'ņ'→'n'	'©'→''	'Å'→'A'
'Ŗ'→'R'	'®'→''	'Ø'→'O'
'ŗ'→'r'	'°'→''	'Æ'→'AE'
'Š'→'S'	'µ'→''	'Ã'→'A'
'š'→'s'	'·'→''	'Õ'→'O'
'Ū'→'U'	'–'→''	'Đ'→'D'
'ū'→'u'	'&mdash'→''	'Ý'→'Y'
'Ą'→'A'	'№'→''	'Ÿ'→'Y'
'ą'→'a'	'Č'→'C'	
'Ć'→'C'	'č'→'c'	
'ć'→'c'	'Š'→'S'	
'Ł'→'L'	'š'→'s'	

Appendix 3.

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

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

'CHEMIE-ING. GRAD.'   'DIPL.-CHEM. DR.'

'DR.DIPL.-CHEM.'   'LANDWIRT'

'DR.-ING. MECH.'   'DR.-ING.'

'-ING. MECH.'   'PROF.DR.'

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

'DIPL.-CHEM.'   '-CHEM.'

'DIPL.-MATH.'   'DR.-MATH.'

'DIPL.-PHYS.'   '-MATH.'

'DIPL.-ING.'   'TECHN.'

'ING.- GRAD'   'DR.-PHYS.'

'ING. GRAD.'   '-PHYS.'

'DIPL.-BIO.'   'DIPL.-'

'IR.-CHEM.'   'PH. D.'

'PROF. DR.'   'DIPL.'

'RER. NAT.'   'PROF.'

'NAT.RER.'   'PH.D.'

'-INFORM.'   '-ING.'

'DIPL-ING'   'CHEM.'

'WIRT.'	'-ING'
'PHYS.'	'ING.'
'PHIL.'	'VET.'
'GRAD.'	'DR.'
'-BIO.'	'DR,'
'MED.'	'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).

2006

CREAP2006-01
Matas, A. (GEAP)**; Raymond, J.Ll.** (GEAP)
"Economic development and changes in car ownership patterns"
(Juny 2006)

CREAP2006-02
Trillas, F. (IEB)**; Montolio, D.** (IEB)**; Duch, N.** (IEB)
"Productive efficiency and regulatory reform: The case of Vehicle Inspection Services"
(Setembre 2006)

CREAP2006-03
Bel, G. (PPRE-IREA)**; Fageda, X.** (PPRE-IREA)
"Factors explaining local privatization: A meta-regression analysis"
(Octubre 2006)

CREAP2006-04
Fernàndez-Villadangos, L. (PPRE-IREA)
"Are two-part tariffs efficient when consumers plan ahead?: An empirical study"
(Octubre 2006)

CREAP2006-05
Artís, M. (AQR-IREA)**; Ramos, R.** (AQR-IREA)**; Suriñach, J.** (AQR-IREA)
"Job losses, outsourcing and relocation: Empirical evidence using microdata"
(Octubre 2006)

CREAP2006-06
Alcañiz, M. (RISC-IREA)**; Costa, A.; Guillén, M.** (RISC-IREA)**; Luna, C.; Rovira, C.**
"Calculation of the variance in surveys of the economic climate"
(Novembre 2006)

CREAP2006-07
Albalate, D. (PPRE-IREA)
"Lowering blood alcohol content levels to save lives: The European Experience"
(Desembre 2006)

CREAP2006-08
Garrido, A. (IEB)**; Arqué, P.** (IEB)
"The choice of banking firm: Are the interest rate a significant criteria?"
(Desembre 2006)

CREAP2006-09
Segarra, A. (GRIT)**; Teruel-Carrizosa, M.** (GRIT)
"Productivity growth and competition in spanish manufacturing firms:
What has happened in recent years?"
(Desembre 2006)

CREAP2006-10
Andonova, V.; Díaz-Serrano, Luis. (CREB)
"Political institutions and the development of telecommunications"
(Desembre 2006)

CREAP2006-11
Raymond, J.L.(GEAP)**; Roig, J.L..** (GEAP)
"Capital humano: un análisis comparativo Catalunya-España"
(Desembre 2006)

CREAP2006-12
Rodríguez, M.(CREB)**; Stoyanova, A.** (CREB)
"Changes in the demand for private medical insurance following a shift in tax incentives"
(Desembre 2006)

CREAP2006-13
Royuela, V. (AQR-IREA)**; Lambiri, D.**; **Biagi**, **B.**
"Economía urbana y calidad de vida. Una revisión del estado del conocimiento en España"
(Desembre 2006)

CREAP2006-14
Camarero, M.; Carrion-i-Silvestre, J.LL. (AQR-IREA).;**Tamarit**, **C.**
"New evidence of the real interest rate parity for OECD countries using panel unit root tests with breaks"
(Desembre 2006)

CREAP2006-15
Karanassou, M.; Sala, H. (GEAP).;**Snower** , **D. J.**
"The macroeconomics of the labor market: Three fundamental views"
(Desembre 2006)

2007

XREAP2007-01
Castany, L (AQR-IREA)**; López-Bazo, E.** (AQR-IREA)**.;Moreno** , **R.** (AQR-IREA)
"Decomposing differences in total factor productivity across firm size"
(Març 2007)

XREAP2007-02
Raymond, J. Ll. (GEAP)**; Roig, J. Ll.** (GEAP)
"Una propuesta de evaluación de las externalidades de capital humano en la empresa"
(Abril 2007)

XREAP2007-03
Durán, J. M. (IEB)**; Esteller, A.** (IEB)
"An empirical analysis of wealth taxation: Equity vs. Tax compliance"
 (Juny 2007)

XREAP2007-04
Matas, A. (GEAP)**; Raymond, J.Ll.** (GEAP)
"Cross-section data, disequilibrium situations and estimated coefficients: evidence from car ownership demand"
 (Juny 2007)

XREAP2007-05
Jofre-Montseny, J. (IEB)**; Solé-Ollé, A.** (IEB)
"Tax differentials and agglomeration economies in intraregional firm location"
 (Juny 2007)

XREAP2007-06
Álvarez-Albelo, C. (CREB)**; Hernández-Martín, R.**
"Explaining high economic growth in small tourism countries with a dynamic general equilibrium model"
 (Juliol 2007)

XREAP2007-07
Duch, N. (IEB)**; Montolio, D.** (IEB); **Mediavilla, M.**
"Evaluating the impact of public subsidies on a firm's performance: a quasi-experimental approach"
 (Juliol 2007)

XREAP2007-08
Segarra-Blasco, A. (GRIT)
"Innovation sources and productivity: a quantile regression analysis"
 (Octubre 2007)

XREAP2007-09
Albalate, D. (PPRE-IREA)
"Shifting death to their Alternatives: The case of Toll Motorways"
 (Octubre 2007)

XREAP2007-10
Segarra-Blasco, A. (GRIT); **Garcia-Quevedo, J.** (IEB); **Teruel-Carrizosa, M.** (GRIT)
"Barriers to innovation and public policy in catalonia"
 (Novembre 2007)

XREAP2007-11
Bel, G. (PPRE-IREA); **Foote, J.**
"Comparison of recent toll road concession transactions in the United States and France"
 (Novembre 2007)

XREAP2007-12
Segarra-Blasco, A. (GRIT);
"Innovation, R&D spillovers and productivity: the role of knowledge-intensive services"
 (Novembre 2007)

XREAP2007-13
Bermúdez Morata, Ll. (RFA-IREA); **Guillén Estany, M.** (RFA-IREA), **Solé Auró, A**. (RFA-IREA)
"Impacto de la inmigración sobre la esperanza de vida en salud y en discapacidad de la población española"
 (Novembre 2007)

XREAP2007-14
Calaeys, P. (AQR-IREA); **Ramos, R.** (AQR-IREA), **Suriñach, J**. (AQR-IREA)
"Fiscal sustainability across government tiers"
 (Desembre 2007)

XREAP2007-15
Sánchez Hugalbe, A. (IEB)
"Influencia de la inmigración en la elección escolar"
 (Desembre 2007)

2008

XREAP2008-01
Durán Weitkamp, C. (GRIT)**; Martín Bofarull, M.** (GRIT) ; **Pablo Martí, F.**
"Economic effects of road accessibility in the Pyrenees: User perspective"
(Gener 2008)

XREAP2008-02
Díaz-Serrano, L.; Stoyanova, A. P. (CREB)
"The Causal Relationship between Individual's Choice Behavior and Self-Reported Satisfaction: the Case of Residential Mobility in the EU"
(Març 2008)

XREAP2008-03
Matas, A. (GEAP)**; Raymond, J. L.** (GEAP)**; Roig, J. L.** (GEAP)
"Car ownership and access to jobs in Spain"
(Abril 2008)

XREAP2008-04
Bel, G. (PPRE-IREA) **; Fageda, X.** (PPRE-IREA)
"Privatization and competition in the delivery of local services: An empirical examination of the dual market hypothesis"
(Abril 2008)

XREAP2008-05
Matas, A. (GEAP); **Raymond, J. L.** (GEAP); **Roig, J. L.** (GEAP)
"**Job accessibility and employment probability**"
(Maig 2008)

XREAP2008-06
Basher, S. A.; Carrión, J. Ll. (AQR-IREA)
Deconstructing Shocks and Persistence in OECD Real Exchange Rates
(Juny 2008)

XREAP2008-07
Sanromá, E. (IEB)**; Ramos, R.** (AQR-IREA); Simón, H.
Portabilidad del capital humano y asimilación de los inmigrantes. Evidencia para España
(Juliol 2008)

XREAP2008-08
Basher, S. A.; Carrión, J. Ll. (AQR-IREA)
Price level convergence, purchasing power parity and multiple structural breaks: An application to US cities
(Juliol 2008)

XREAP2008-09
Bermúdez, Ll. (RFA-IREA)
A priori ratemaking using bivariate poisson regression models
(Juliol 2008)

XREAP2008-10
Solé-Ollé, A. (IEB), **Hortas Rico, M.** (IEB)
Does urban sprawl increase the costs of providing local public services? Evidence from Spanish municipalities
(Novembre 2008)

XREAP2008-11
Teruel-Carrizosa, M. (GRIT), **Segarra-Blasco, A.** (GRIT)
Immigration and Firm Growth: Evidence from Spanish cities
(Novembre 2008)

XREAP2008-12
Duch-Brown, N. (IEB), **García-Quevedo, J.** (IEB), **Montolio, D.** (IEB)
Assessing the assignation of public subsidies: Do the experts choose the most efficient R&D projects?
(Novembre 2008)

XREAP2008-13
Bilotkach, V., **Fageda, X.** (PPRE-IREA), **Flores-Fillol, R.**
Scheduled service versus personal transportation: the role of distance
(Desembre 2008)

XREAP2008-14
Albalate, D. (PPRE-IREA), **Gel, G.** (PPRE-IREA)
Tourism and urban transport: Holding demand pressure under supply constraints
 (Desembre 2008)

2009

XREAP2009-01
Calonge, S. (CREB)**; Tejada, O.**
"A theoretical and practical study on linear reforms of dual taxes"
(Febrer 2009)

XREAP2009-02
Albalate, D. (PPRE-IREA)**; Fernández-Villadangos, L.** (PPRE-IREA)
"Exploring Determinants of Urban Motorcycle Accident Severity: The Case of Barcelona"
(Març 2009)

XREAP2009-03
Borrell, J. R. (PPRE-IREA)**; Fernández-Villadangos, L.** (PPRE-IREA)
"Assessing excess profits from different entry regulations"
(Abril 2009)

XREAP2009-04
Sanromá, E. (IEB)**; Ramos, R.** (AQR-IREA)**, Simon, H.**
"Los salarios de los inmigrantes en el mercado de trabajo español. ¿Importa el origen del capital humano?"
(Abril 2009)

XREAP2009-05
Jiménez, J. L.; Perdiguero, J. (PPRE-IREA)
"(No)competition in the Spanish retailing gasoline market: a variance filter approach"
(Maig 2009)

XREAP2009-06
Álvarez-Albelo,C. D. (CREB)**, Manresa, A.** (CREB)**, Pigem-Vigo, M.** (CREB)
"International trade as the sole engine of growth for an economy"
(Juny 2009)

XREAP2009-07
Callejón, M. (PPRE-IREA)**, Ortún V, M.**
"The Black Box of Business Dynamics"
(Setembre 2009)

XREAP2009-08
Lucena, A. (CREB)
"The antecedents and innovation consequences of organizational search: empirical evidence for Spain"
(Octubre 2009)

XREAP2009-09
Domènech Campmajó, L. (PPRE-IREA)
"Competition between TV Platforms"
(Octubre 2009)

XREAP2009-10
Solé-Auró, A. (RFA-IREA),**Guillén, M.** (RFA-IREA), **Crimmins, E. M.**
"Health care utilization among immigrants and native-born populations in 11 European countries. Results from the Survey of Health, Ageing and Retirement in Europe"
(Octubre 2009)

XREAP2009-11
Segarra, A. (GRIT)**, Teruel, M.** (GRIT)
"Small firms, growth and financial constraints"
(Octubre 2009)

XREAP2009-12
Matas, A. (GEAP)**, Raymond, J.Ll.** (GEAP), **Ruiz, A.** (GEAP)
"Traffic forecasts under uncertainty and capacity constraints"
(Novembre 2009)

XREAP2009-13
Sole-Ollé, A. (IEB)
"Inter-regional redistribution through infrastructure investment: tactical or programmatic?"
(Novembre 2009)

XREAP2009-14
Del Barrio-Castro, T., **García-Quevedo, J.** (IEB)
"The determinants of university patenting: Do incentives matter?"
(Novembre 2009)

XREAP2009-15
Ramos, R. (AQR-IREA), **Suriñach, J.** (AQR-IREA), **Artís, M.** (AQR-IREA)
"Human capital spillovers, productivity and regional convergence in Spain"
(Novembre 2009)

XREAP2009-16
Álvarez-Albelo, C. D. (CREB), **Hernández-Martín, R.**
"The commons and anti-commons problems in the tourism economy"
(Desembre 2009)

2010

XREAP2010-01
García-López, M. A. (GEAP)
"The Accessibility City. When Transport Infrastructure Matters in Urban Spatial Structure"
(Febrer 2010)

XREAP2010-02
García-Quevedo, J. (IEB), **Mas-Verdú, F.** (IEB), **Polo-Otero, J.** (IEB)
"Which firms want PhDs? The effect of the university-industry relationship on the PhD labour market"
(Març 2010)

XREAP2010-03
Pitt, D., **Guillén, M.** (RFA-IREA)
"An introduction to parametric and non-parametric models for bivariate positive insurance claim severity distributions"
(Març 2010)

XREAP2010-04
Bermúdez, Ll. (RFA-IREA), **Karlis, D.**
"Modelling dependence in a ratemaking procedure with multivariate Poisson regression models"
(Abril 2010)

XREAP2010-05
Di Paolo, A. (IEB)
"Parental education and family characteristics: educational opportunities across cohorts in Italy and Spain"
(Maig 2010)

XREAP2010-06
Simón, H. (IEB), **Ramos, R.** (AQR-IREA), **Sanromá, E.** (IEB)
"Movilidad ocupacional de los inmigrantes en una economía de bajas cualificaciones. El caso de España"
(Juny 2010)

XREAP2010-07
Di Paolo, A. (GEAP & IEB), **Raymond, J. Ll.** (GEAP & IEB)
"Language knowledge and earnings in Catalonia"
(Juliol 2010)

XREAP2010-08
Bolancé, C. (RFA-IREA), **Alemany, R.** (RFA-IREA), **Guillén, M.** (RFA-IREA)
"Prediction of the economic cost of individual long-term care in the Spanish population"
(Setembre 2010)

XREAP2010-09
Di Paolo, A. (GEAP & IEB)
"Knowledge of catalan, public/private sector choice and earnings: Evidence from a double sample selection model"
(Setembre 2010)

XREAP2010-10
Coad, A., **Segarra, A.** (GRIT), **Teruel, M.** (GRIT)
"Like milk or wine: Does firm performance improve with age?"
(Setembre 2010)

XREAP2010-11
Di Paolo, A. (GEAP & IEB), **Raymond, J. Ll.** (GEAP & IEB), **Calero, J.** (IEB)
"Exploring educational mobility in Europe"
(Octubre 2010)

XREAP2010-12
Borrell, A. (GiM-IREA), **Fernández-Villadangos, L.** (GiM-IREA)
"Clustering or scattering: the underlying reason for regulating distance among retail outlets"
(Desembre 2010)

XREAP2010-13
Di Paolo, A. (GEAP & IEB)
"School composition effects in Spain"
(Desembre 2010)

XREAP2010-14
Fageda, X. (GiM-IREA), **Flores-Fillol, R.**
"Technology, Business Models and Network Structure in the Airline Industry"
(Desembre 2010)

XREAP2010-15
Albalate, D. (GiM-IREA), **Bel, G.** (GiM-IREA), **Fageda, X.** (GiM-IREA)
"Is it Redistribution or Centralization? On the Determinants of Government Investment in Infrastructure"
(Desembre 2010)

XREAP2010-16
Oppedisano, V., **Turati, G.**
"What are the causes of educational inequalities and of their evolution over time in Europe? Evidence from PISA"
(Desembre 2010)

XREAP2010-17
Canova, L., **Vaglio, A.**
"Why do educated mothers matter? A model of parental help"
(Desembre 2010)

2011

XREAP2011-01
Fageda, X. (GiM-IREA), **Perdiguero, J.** (GiM-IREA)
"An empirical analysis of a merger between a network and low-cost airlines"
(Maig 2011)

XREAP2011-02
Moreno-Torres, I. (ACCO, CRES & GiM-IREA)
"What if there was a stronger pharmaceutical price competition in Spain? When regulation has a similar effect to collusion"
(Maig 2011)

XREAP2011-03
Miguélez, E. (AQR-IREA); Gómez-Miguélez, I.
"Singling out individual inventors from patent data"
(Maig 2011)

xreap@pcb.ub.es