Research Policy 32 (2003) 69–87

research policy

www.elsevier.com/locate/econbase

Knowledge-relatedness in firm technological diversification

Stefano Breschi, Francesco Lissoni*, Franco Malerba

Cespri-Università Bocconi, via Sarfatti 25, 20136 Milan, Italy

Received 1 August 2000; received in revised form 29 January 2001; accepted 7 December 2001

Abstract

This paper claims that knowledge-relatedness is a key factor in affecting firms' technological diversification. The hypothesis is tested that firms extend the range of their innovative activities in a non-random way. Specifically, we test the extent to which firms diversify their innovative activities across related technological fields, i.e. fields that share a common knowledge base and rely upon common heuristics and scientific principles. The paper proposes an original measure of knowledge-relatedness, using co-classification codes contained in patent documents, and examines the patterns of technological diversification of the whole population of firms from the United States, Italy, France, UK, Germany, and Japan patenting to the European Patent Office from 1982 to 1993. Robust evidence is found that knowledge-relatedness is a major feature of firms' innovative activities.

JEL classification: C81; L20; 031

Keywords: Diversification; Relatedness; Patents; Co-classification

1. Introduction

This paper is about the extent and the nature of the range of firms' innovative activities and the role of knowledge-relatedness in affecting firms' technological diversification.

Most of the times in their innovative activities firms span over more than one technology, i.e. they are 'technologically diversified'. The literature on innovation and technical change has evidenced some robust stylised facts about firms' technological diversification. First, technological diversification is usually greater than product diversification. Firms have to manage a wide number of technologies in order to develop and produce products and services. Thus, most firms could be labelled multi-technology corporations, even if they are specialised in just one line of business (Granstrand, 1998). Second, most of the times technological diversification anticipates product and market diversification (Pavitt, 1998). This is so because technological exploration in a wide range of technologies is a prerequisite for production. Third, the profile of technological diversification of firms is rather stable. It changes slowly over time as a consequence of the inertia of specialisation, incremental changes in knowledge production and modifications in firms' competencies (Cantwell and Andersen, 1996). Fourth, the profile of technological diversification differs across large firms, as a consequence of the history of the corporations, initial conditions, the specialisation of the companies, the market incentives and the specific institutional setting in which companies are embedded. Fifth, the profile of technological diversification is very similar among large firms producing

* Corresponding author. Tel.: +39-2-58363376; fax: +39-2-58363399.
E-mail address: lissoni@ing.unibs.it (F. Lissoni).

0048-7333/02/$ – see front matter
PII: S0048-7333(02)00004-5

similar products, particularly in high technology and technology-based industries (Patel and Pavitt, 1995).

Based upon such evidence, a question that has recently gained the attention of several scholars in the field of industrial organisation and technical change concerns the nature and the determinants of firms' productive and technological diversification. There is emerging evidence that the range of firms' technological and productive activities, far from being driven by short-term quest for profits, follows some 'purposiveness' (Scott, 1993). Relatedly, it has been shown that firms exhibit some coherence in the technological and productive activities they are engaged in (MacDonald, 1985; Teece et al., 1994). However, the notion, determinants and measurement of the coherence of firms are still to be fully developed. In this respect, much work has still to be done at both the conceptual and the empirical levels.

This paper goes in this direction, by exploring the extent and the nature of firms' technological diversification and by linking it to a key factor: knowledge-relatedness. In order to do that, the paper first tests whether firms' technological activities are not distributed randomly across technological fields (*random hypothesis*) and confirms Teece's et al. (1994) seminal findings.

In addition, the paper tests (more directly than ever before) the *knowledge-relatedness hypothesis*, claiming that firms follow a coherent pattern of technological diversification, which clusters around groups of technologies that share a common or complementary knowledge base, rely upon common scientific principles or have similar heuristics of search.

The concept of knowledge-relatedness is very broad and encompasses several dimensions of knowledge. They can be grouped into three categories: proximity, commonality and complementarity.

a. *Knowledge proximity*

Knowledge proximity may be the result of learning processes, which could be conventionally considered either unintended (*learning spillovers*, as emphasised by the traditional literature on productivity growth) or intended (*local learning*).

Learning spillovers relate to the fact that knowledge coming from activities in one technology may spill over to other technologies that are related in various ways. A firm active in one technology may find itself engaged into another technology because there are 'knowledge externalities and spillovers' (Griliches, 1979; Henderson and Cockburn, 1996). According to this explanation, the presence of a firm in more than one technological field is the unintended outcome of its innovative activities.

In particular, *local learning* relate to the fact that firms may focus (often intentionally) on new technologies that are rather similar to the ones the firm is currently developing in terms of knowledge base (such as scientific fields, search procedures—such as R&D and learning by doing—sources of knowledge-such as universities, public research institutes, users, suppliers, and so on). Thus, firms' innovative search takes place in the neighbourhood of the technologies currently developed and innovative activities proceed incrementally (Atkinson and Stiglitz, 1969; David, 1975; Malerba, 1992; Antonelli, 1995). According to this explanation, because of uncertainty and change, firms are not able to find the best and most promising or profitable technologies. Rather, they are boundedly rational actors that focus on technological domains, which present similarity in problem solving and knowledge bases (Nelson and Winter, 1982; Dosi, 1997). In this case, the presence of a firm in more than one technological field has to be interpreted as the intended outcome of its innovative activities.

b. *Knowledge commonalities*

Firms' innovative activities may span over more than one technologies because the same type of knowledge is used in more than one technology. Therefore, firms have economies of scope in the 'use of one piece of knowledge'. Penrose (1959) and Teece (1982) have discussed the role of common resources in relation to firms' business diversification. Here, the same type of knowledge is an input for the innovative activity in two technological fields.

c. *Knowledge complementarities*

Complementarities refer to *complementary knowledge and technologies*. In order to introduce new products and new processes, firms may have to master together more than one technology. In this case, the relatedness among two or more technologies is not due to their similarities, but to their differences and to the need to use them together. The relevance of complementarities have been emphasised by a wide range of authors with different focuses (Milgrom and Roberts, 1990; Scott, 1993, Pavitt, 1998).

Generic technologies are an extension of the previous explanation. When a complementary technology is used in a wide range of other technologies for the development of different products and processes, it may become a generic technology (Arora and Gambardella, 1994; Bresnahan and Trajtenberg, 1995). According to this view, a generic technology is a highly pervasive complementary technology.

In sum, technological relatedness is due to learning processes (unintended in terms of spillovers and intended in terms of local learning) and knowledge links (due to the scope, complementarity or generic nature of knowledge).

Note that to the explanations of technological diversification based on knowledge proximity and knowledge commonality, more traditional ones related to sunk costs and switching costs may be added. Firms may find difficult to move away from current technologies because of organisation and R&D costs and the presence of R&D equipment and human capital. These are factors that lock firms in the neighbourhood of the technologies they are currently using. Due to high switching costs, firms will move locally into 'related technologies' rather than into unrelated ones. This cost side aspect, however, will not be discussed in this paper.[1]

This paper will proceed in two steps. First, in Sections 2 and 3, it will assess empirically the extent and direction of technological diversification for the whole population of firms involved in innovative activities. The aim is to provide a descriptive analysis of the technological diversification of the whole population of innovators, large as well as small, and not just of the largest ones. Innovators will be examined in the six major-advanced countries—Unites States, Japan, Germany, France, UK and Italy—for the period 1982–1993. Second, in Sections 4–6, it will test the hypothesis that knowledge-relatedness is a major driver of technological diversification against the hypothesis that technological diversification proceeds randomly. If the relatedness hypothesis holds, it means that firms diversify technologically along certain directions that depend on the links and distance among technological fields.

[1] We will not discuss either the presence of firms in technologies that may look promising (opening up windows), which is common in high technologies.

2. Data sources

This paper is based upon the EPO-CESPRI database. The data set contains all patent applications to the European Patent Office (EPO) from 1978 (EPO's first year of activity) to 1993, by firms, institutions and individuals of all countries seeking legal protection for their innovations in any of the 18 countries adhering to the Munich Convention which established the EPO. In addition to measuring innovation (with all the strengths and weaknesses and the methodological problems associated to this; see Griliches, 1991), *patent applications* are a very good indicator of firms' technological competencies. The fact that a firm applies for a patent in a given technological field means that such a firm is at, or close to, the technological frontier and has advanced technological competencies in that field.

CESPRI has developed the basic EPO data in the following two different ways:

- Patent applications have been processed *at the firm level*, for all firms from France, Germany, Italy, Japan, UK and United States. For each patent document, therefore, the EPO-CESPRI database contains information about: (a) the name and the address of the patenting firm; (b) the date of filing to the EPO; (c) the technological field which was assigned by patent examiners.[2]

[2] The EPO-CESPRI dataset contains 17,394 patents and 4802 firms for Italy, 124,626 patents and 10,459 firms for Germany, 39,582 patents and 7121 firms for the UK, 51,690 patents and 6835 firms for France, 164,790 patents and 14,395 firms for the US and 113,629 patents and 5025 firms for Japan. Firms that are part of business groups have been treated as individual companies. In case of co-patenting, each co-patentee has been credited the patent. Individual inventors have been excluded from the dataset. Since individual inventors are mostly self-employed and owners of small independent firms, their exclusion from the data set could underestimate the contribution of smaller companies to the innovative activities. However, the share of total patent applications held by private individuals in the dataset is rather small (generally, <3% of total patent applications). Finally, one must note that since the EPO is located in Germany, German firms are over-represented

Table 1
Technology classification (30 fields) based on the IPC

1. Electrical engineering	16. Chemical engineering
2. Audiovisual technology	17. Surface technology
3. Telecommunications	18. Materials processing
4. Information technology	19. Thermal processes
5. Semiconductors	20. Environmental technology
6. Optics	21. Machine tools
7. Control technology	22. Engines
8. Medical technology	23. Mechanical elements
9. Organic chemistry	24. Handling
10. Polymers	25. Food processing
11. Pharmaceutics	26. Transport
12. Biotechnology	27. Nuclear engineering
13. Materials	28. Space technology
14. Food chemistry	29. Consumer goods
15. Basic materials chemistry	30. Civil engineering

- Patent applications to the EPO have also been used to measure the *knowledge-relatedness* between different technological fields. In this case, *all* patent applications to the EPO have been considered, i.e. not only those coming from the above-mentioned countries, but also those coming from any other country in the world.

For both types of elaboration a common technological classification has been used. More specifically, we adopted a technology-oriented classification that distinguishes 30 different fields of technology based on the International Patent Classification (IPC). This classification has been elaborated jointly by Fraunhofer Gesellschaft-ISI (Karlsruhe), *Institut National de la Propriété Industrielle* (INPI, Paris) and *Observatoire des Sciences and des Techniques* (OST, Paris) and it is reported in the Table 1.[3]

Although the EPO database starts in 1978, our analysis will cover only the period from 1982 to 1993. In fact, the sample of patent applications from 1978 to 1982 could be biased by the fact that, when the EPO started its activity and immediately afterward, only the largest and established firms (and especially the German ones) were likely to know it well enough to apply for patents.

In the following section, we start exploring the database at the firm level and offer some preliminary evidence on the relevance of the phenomenon of technological diversification, as well as a collection of stylised facts that suggest the existence of a close connection between technological diversification and persistence.

3. Technological diversification: a few stylised facts

Using the EPO-CESPRI database, we can identify the following two basic types of innovative firms:[4]

(i) *Diversified innovators*, i.e. those firms that took patents in more than one technological field over the period 1982–1993.
(ii) *Specialised innovators*, i.e. those firms that took patents in only one technological field over the period 1982–1993.[5]

For descriptive purposes, we can also identify a third category of innovative firms, which comprises both diversified and specialised innovators:

(iii) *Persistent innovators*, i.e. those firms that took patents in all the three subperiods 1982–1985, 1986–1989 and 1990–1993, not necessarily in the same technological field.[6]

in the sample. However, because the focus in this paper is not on absolute technological performance, but on the patterns of technological diversification, we think that this does not create too serious a distortion in our results.

[3] All patent documents are indeed assigned by patent examiners of the EPO at least one classification code of the IPC. The IPC is an internationally agreed, non-overlapping and comprehensive patent classification system. Currently, the IPC refers to almost 60,000 individual codes (12 digits) and it may be used at different hierarchical levels (WIPO, 1994). The concordance of IPC to our 30-field classification is available on request.

[4] In what follows, for sake of simplicity, we will use the terms ‘innovators’ and ‘innovative firms’ instead of ‘patenting firms’. Note that this is a very restrictive use of those terms. A firm may be quite innovative even without patenting its innovations.

[5] Please note that the benchmarking used here to define diversified versus specialised innovators is rather narrow. However, it may be considered a useful approximation for descriptive purposes.

[6] The definition of ‘innovation persistence’ reported in the text is admittedly crude. Various authors have recently measured in a more accurate way the extent to which firms persistently innovate (Geroski et al., 1997; Malerba et al., 1997; Cefis, 1996). However, since the focus of this paper is not on innovation persistence, we think that such definition is acceptable for descriptive purposes.

Table 2
Relative weight of diversified vs. specialised innovators (six countries; percentage values, 1982–1993)

	Diversified innovators	Specialised innovators	Total	No. of observations
Share of *firms*	30.2	69.8	100	44634
Share of *patents*	89.5	10.5	100	451772

Source: EPO-CESPRI database.

Based upon these definitions, the following few stylised facts on technological diversification emerge from the analysis of the data:

(a) *Most patenting firms are specialised in only one field of technology, but they are small innovators*. Table 2, which reports the relative importance of specialised versus diversified innovators, indicates that specialised innovators represent the large majority (almost 70%) of all patenting firms. However, diversified innovators weight disproportionately in terms of patents, accounting for almost 90% of all patent applications from the six countries examined here in the period 1982–1993.

(b) *Most diversified innovators are active in a low number of technological fields, and are relatively small innovators*. For the six countries pooled together, Fig. 1 reports the distribution of diversified innovators according to the number of different technological fields in which they applied for patents in the whole period 1982–1993. More than 50% of all diversified innovators are present in just two technological fields, and around 20% in three. Fig. 2 reports instead the distribution of patents held by diversified innovators according to the number of technological fields in which those innovators applied for a patent in the whole period 1982–1993. The distribution shows rather clearly a U-shaped pattern. Firms diversified in two and three technological fields hold, respectively, only around 7.5 and 4.5% of all patents held by diversified innovators in the period 1982–1993. These results taken together suggest that while most diversified firms are present in

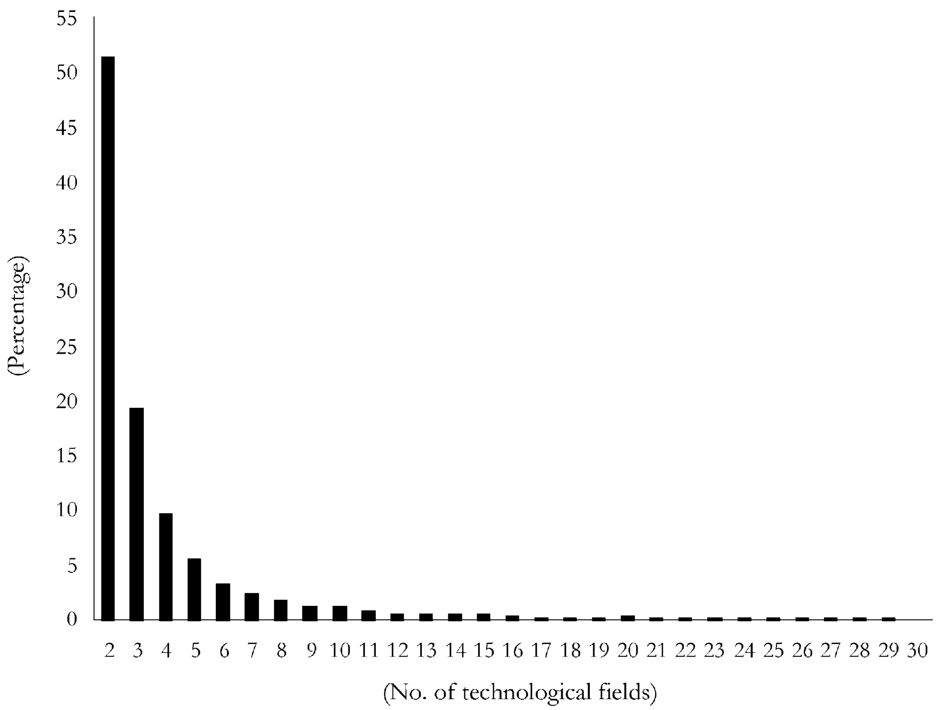


Fig. 1. Distribution of diversified innovators according to the number of technological fields in which they filed for patents (percentage values, 1982–1993).

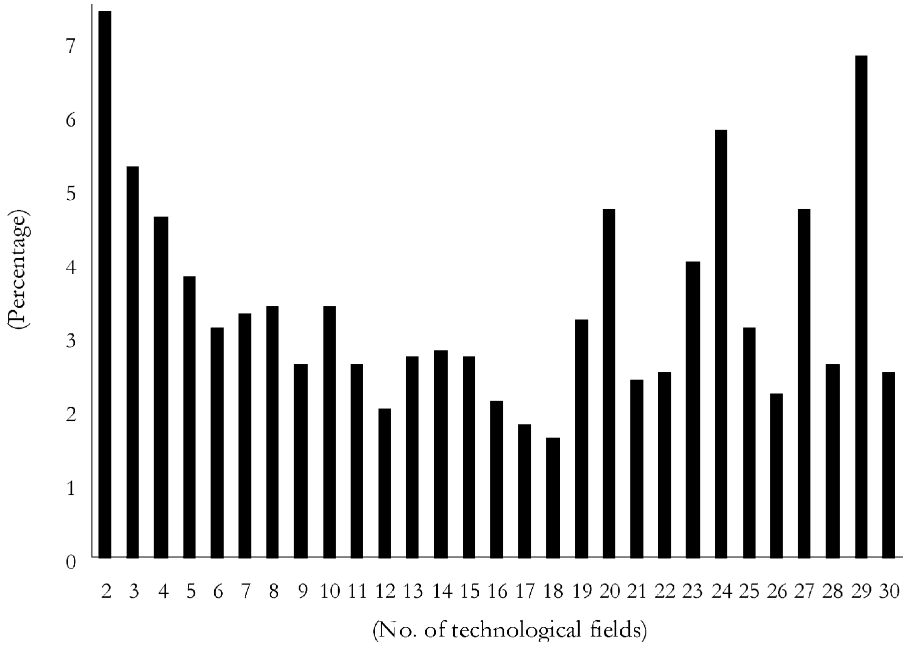


Fig. 2. Distribution of total patents filed by diversified innovators according to the number of technological fields in which they later filed for patents (percentage values, 1982–1993).

a relatively low number of technological fields (e.g. slightly >70 and 80% of them hold patents in less than three and four technological fields, respectively), they are also relatively small innovators (e.g. only around 12 and 17% of all patents are accounted for by firms diversified in less than three and four technological fields, respectively).

(c) *Very few firms are diversified in most technological fields, and these are the largest innovators*. Figs. 1 and 2 suggest that very few firms are present in most technological fields (e.g. only six firms patented in all the 30 technological fields in the period 1982–1993, and only seven patented in 29 and 28 technological fields, respectively). However, such firms are very large innovators. For example, firms diversified in >23 technological fields account for almost 30% of all patent applications filed to the EPO by diversified innovators in the period 1982–1993.

(d) *Most persistent innovators are diversified innovators*. Table 3 reports the share of diversified innovators among persistent ones. Diversified innovators represent >88% of all persistent innovators. The share of diversified innovators is even higher in terms of patents, accounting for almost 99% of all patent applications by persistent innovators in the period 1982–1993. This result, thus, suggests that both in terms of firms and patents, diversified innovators constitute the large majority of persistent innovators.

Summing up, technological diversification is a widespread phenomenon in Europe, the United States and Japan. Generally speaking, only occasional innovators are not diversified. On the contrary, most persistent innovators are diversified in at least two technologies. This suggests that being technologically

Table 3
Share of diversified innovators among persistent innovators (six countries; percentage values, 1982–1993)

	Diversified innovators	Specialised innovators	Total	No. of observations
Share of *firms*	88.4	11.6	100	4775
Share of *patents*	98.8	1.2	100	345819

Source: EPO-CESPRI database.

diversified (i.e. being able to master and use different technologies) may represent a necessary requisite to survive and grow as an innovator, and/or vice versa. In the following section, we will start to examine the nature of firms' technological diversification by analysing to what extent it follows a coherent pattern. In particular, we will test the hypothesis that firms distribute their innovative activities across technological fields in a random way.

4. A test of randomness in patterns of firms' technological diversification

A major claim of this paper is that, whatever the reason why firms have expertise in and choose to master different technological fields, they should do it in a relatively 'coherent' way. In this respect, a firm can be said to have a coherent portfolio of innovative activities, when the technological fields in which it is engaged are *related*, in the sense that they share some knowledge or scientific principles and/or enter as inputs into the production of specific foods or services. By contrast, a firm fails to exhibit a coherent pattern of technological diversification when its innovative activities are randomly distributed across technological fields (Teece et al., 1994). In this section, we show that firms' technological diversification can be hardly seen as a purely random phenomenon, i.e. one driven by random outcomes in firms' innovative activities (*random hypothesis*). To this purpose, we carry out a test proposed by Engelsman and van Raan (1991) and Teece et al. (1994).

Let us indicate with T the total number of firms having patented at the EPO in 1982–1993 *in two or more technological fields*.[7] Let $G_{it} = 1$ if firm t patented in technological field i and $G_{it} = 0$ otherwise. The total number of firms having patented in technological field i in 1982–1993 is therefore given by: $R_i = \Sigma_t G_{it}$. Using this notation, we can also indicate the number of firms that have patented, i.e. are active in both technological fields i and j as follows: $O_{ij} = \Sigma_t G_{it} G_{jt}$. By applying the latter to all possible pairs of technological fields we obtain a square (30 × 30) symmetrical matrix $\boldsymbol{\Omega}$, whose generic cell O_{ij} reports the number of firms that in 1982–1993 were active in both technological fields i and j. Matrices like $\boldsymbol{\Omega}$ were calculated for the whole population of diversified innovators, and for particular subsets of them, such as the sample of innovators which diversify in only two technological fields, up to three fields, up to four fields and so on.

A test of randomness can thus be performed by comparing the observed value of O_{ij} with the value that would be expected under the hypothesis that technological diversification is random. More particularly, let us assume that in a population of T innovative firms, R_i firms possess the characteristic of being active in technological field i. This implies, of course, that $(T - R_i)$ firms do not possess such characteristic. Now, an independent sample (without replacement) of size R_j of firms is drawn from the population of T innovative firms: these firms are assigned activities in technological field j. Given this experiment, the probability of obtaining exactly x firms that are active in both technological fields i and j is distributed according to a hypergeometric random variable, with population T, i-field members R_i, and sample size R_j:

$$P[X_{ij} = x] = \frac{\binom{R_i}{x}\binom{T - R_i}{R_j - x}}{\binom{T}{R_j}} \tag{1}$$

The mean and the variance of X_{ij} are, respectively,

$$\mu_{ij} = E(X_{ij}) = \frac{R_i R_j}{T} \tag{2}$$

$$\sigma_{ij}^2 = \mu_{ij}\left(1 - \frac{R_i}{T}\right)\left(\frac{T}{T-1}\right) \tag{3}$$

A test of randomness in firms' technological diversification can thus be based upon the following statistic:

$$r_{ij} = \frac{O_{ij} - \mu_{ij}}{\sigma_{ij}}, \tag{4}$$

where O_{ij} is the value of the generic cell of matrix $\boldsymbol{\Omega}$; μ_{ij} and σ_{ij} are respectively the mean and variance of the hypergeometric distribution that we would expect to obtain under the random hypothesis. This statistic measures, therefore, the extent to which the observed

[7] The analysis that follows is based upon diversified innovators from six countries: Italy, France, Germany, UK, Japan and United States. Moreover, since specialised innovators shed no light on the issue examined here, they are omitted from the analysis.

Table 4
Test of randomness in firms' technological diversification (r_{ij} index) (number and percentage of cases with significant P-values[a], six countries, 1982–1993)

No. of fields[b]	No. of firms[c]	Positive[d]		Negative[e]		All		Min	Max	Mean	S.D.
		#	%	#	%	#	%				
Up to 2	6986	2	4.60	211	48.50	231	53.10	−10.35	32.28	−2.58	3.67
Up to 3	9507	3	8.96	240	55.17	279	64.13	−12.77	42.18	−2.73	5.22
Up to 4	10804	5	11.95	234	53.79	286	65.74	−13.41	46.34	−2.50	6.33
Up to 5	11548	6	14.71	232	53.33	296	68.04	−13.44	49.82	−2.11	7.13
Up to 6	11977	7	16.78	222	51.03	295	67.81	−14.07	52.33	−1.66	7.76
Up to 30	13466	36	84.82	15	3.45	384	88.27	−11.14	61.68	11.63	10.29

Source: EPO-CESPRI database.
[a] Number of technological fields in which diversified innovators have patented.
[b] Total number of patenting firms in each frequency class.
[c] Number and percentage of cells with a positive and statistically significant index of relatedness.
[d] Number and percentage of cells with a negative and statistically significant index of relatedness.
[e] Total number of cases with a significant P-value.

association between two technological fields exceeds what we expected if firms were assigned to technological fields randomly. If the actual number of firms diversified in technological fields i and j (i.e. O_{ij}) greatly exceeds the expected number μ_{ij}, then there must be a strong (non random) relationship between the two technological fields. If, on the contrary, r_{ij} takes a negative value, this means that O_{ij} is even lower than the number we would observe if firms were to choose their technological fields randomly.

A major advantage of the r_{ij} index is that it is possible to calculate the P-values for each element r_{ij} under the null hypothesis of independence between technological fields and therefore to evaluate the statistical significance of the observed relationship among them. Using these properties, we have calculated the r_{ij} index and the associated P-values for various subsets of diversified innovators, thus taking into account the possibility that the degree of association among technological fields could vary with the number of fields in which innovators are diversified. Table 4 reports the main results of this analysis by showing summary statistics on the absolute number and percentage of cases (i.e. pairs of technological fields) with statistically significant P-values (10% level). The results reject the random hypothesis. In fact, the observed diversification patterns appear to be significantly more linked (or not linked) than it would happen under the random hypothesis. The percentage of cases with non-random (negative or positive) relationships among technological fields is always well above 50% and it increases by including firms with a wider portfolio of innovative activities. Moreover, as one would expect, the percentage of technological fields strongly (i.e. positively) associated sharply increases when we consider highly diversified firms.[8]

On the basis of these results, one could be tempted to conclude that there is a tendency for firms to innovate in clusters of *related* technologies, in the sense defined above. In our view, this conclusion is not warranted without making further assumptions. For example, Teece et al. (1994) invoke the *survivor principle*, according to which economic competition will lead to the disappearance of relatively inefficient organisational forms, and assume that (industrial) activities which are more frequently combined within the same firm must therefore be more related. Our argument is that this kind of inference should be based only upon an objective and direct measure of knowledge or technological relatedness among industries or technological fields. In the following section, we will propose a measure of knowledge-relatedness among technological fields that allow to test directly to what extent firms' technological diversification takes place across *related* technological fields.

[8] Please note that the analysis carried out in this section does not tell anything about the *overall* degree of diversification of firms. In a related paper (Breschi et al., 2002), we measure the overall degree of firms' technological coherence by adopting two indices of weighted-average-relatedness originally proposed by Teece et al. (1994).

5. Measuring knowledge-relatedness

In recent years there have been various attempts to conceptualise relatedness among technological fields and to find appropriate measures for it.[9] For long, the most influential approach has been Scherer's (1982) method for measuring inter-industry technology flows.[10] This was based upon a classification of R&D outlays by industry of origin and industry of use of the resulting products and processes. According to this method, two industries are considered *close* to each other if a rather high share of the R&D performed in one sector is actually embodied and used in the other one. More recently, an alternative approach has been proposed by Jaffe (1986, 1989), who measured technological relatedness among a sample of US firms by looking at the distribution of their patents over 49 technology fields (each field representing a collection of 12-digit IPC codes). In particular, Jaffe employed the so-called *cosine index* to measure the correlation between a number of vectors representing the distribution of firms' patents over the various fields.[11]

Still more recently, this issue has been tackled upon by bibliometrics. According to this approach, the relatedness between fields of technology can be measured analysing the co-occurrence of classification codes assigned to individual patent documents (Engelsman and van Raan, 1991, 1992).[12] As described above, all patent documents are classified at least by one (*main* or *primary*) classification code of the International Patent Classification (IPC), but usually more classification codes (*secondary* or *supplementary*) are assigned to the documents by the patent examiners of the issuing patent offices. The assumption which is made is that the frequency by which two classification codes are jointly assigned to the same patent document can be interpreted as a sign of the strength of the knowledge relationship, in terms of knowledge links and spillovers, between the technological fields which the codes refer to, i.e. as an inverse measure of the distance between the knowledge bases of the two fields.[13] Verspagen (1997) has recently suggested a method which is based upon the distinction between the *main* classification code assigned to a patent document, and the *supplementary* classification codes that the examiners of the issuing patent office usually add to it to specify in detail the technical content of the novelty claim. According to Verspagen, the main code refers to the object of claimed and appropriable knowledge, while the supplementary code refers to some non-appropriable additional knowledge, i.e. knowledge that is not new and upon which no discovery claim is made. Following this distinction, Verspagen (1997) assumes that the main classification code "*provides a good proxy of the producing sector of knowledge and that the listed supplementary IPC codes (taken as partially unintended "by-products" of the main goal of the invention) give an indication for technology spillovers to other industrial sectors*."

In this paper, we also make use of the different classification codes assigned to patents, but take a

---

[9] In a parallel way, a great deal of work has been also devoted to measure corporate diversification and inter-industry relatedness (Caves, 1981; Lemelin, 1982).

[10] Notably, this approach has been recently adopted to build the so-called 'Yale-matrix' from Canadian Patent Office data (Putnam and Evenson, 1994).

[11] It is worth pointing out that efforts to conceptualise and measure knowledge or technological relatedness among firms or technological fields have been intimately related to the attempts to capture and assess the impact of spillover phenomena (Jaffe, 1986; Griliches, 1995; Grupp, 1996). A thorough discussion of this issue is, however, beyond the scope of this paper.

[12] For Engelsman and van Raan (1991), this approach produces meaningful results particularly at a 'macro' level, i.e. for mapping the entire domain of technology. In order to map relationships at the meso and micro levels (i.e. for individual technology fields or combinations of related fields of technology), they suggest to adopt the so-called *co-word* analysis, i.e. analysis based on the co-occurrence of all indexed keywords contained in patent titles and patent abstracts.

[13] A problem with this interpretation has to do with the fact that classification codes are assigned to patents for 'search' purposes. A related criticism against the use of classification codes to measure knowledge relatedness is that, since they are assigned by patent examiners, they do not necessarily reflect the way firms actually perceive knowledge relatedness among technological fields. A possible way to answer these two criticisms is as follows. IPC codes are used by patent examiners to classify all relevant technical features of a patent application (WIPO, 1994). If different technical aspects are addressed, multiple classification codes are necessary. Of course, the codes are also used for search purposes. So patent examiners scan other documents with the same code for identifying documents with similar features. If a document has several codes, a broader area of search has to be covered. Therefore, a frequent co-classification of two codes, indeed, indicates a technological or knowledge relatedness (since technology includes technical knowledge). We are indebted to Ulrich Schmoch for these observations.

different position. Contrary to Verspagen we make no assumption about the meaning of the main classification codes as opposed to the supplementary ones. As explained by Hinze et al. (1997), and confirmed by a number of WIPO documents,[14] the two kinds of codes cannot be used to distinguish between knowledge-producing and knowledge-incorporating fields. In fact, although the main classification code describes the central characteristics of the main claim of the patent, the supplementary codes indicate further features of the main claim as well as of the remaining claims of the patent, i.e. they also refer to knowledge creation. Therefore, *nothing can be said about the direction of knowledge flows*. Following these remarks, we do not distinguish between main and supplementary codes, and weigh equally all classification codes assigned to a patent document.[15]

The procedure we followed to build a measure of knowledge-relatedness was originally proposed by Engelsman and van Raan (1992). Formally, let us indicate with M the universe being studied, e.g. all patent applications to the EPO in a certain period of time. Each of the M patent applications has been assigned by patent examiners to one or more classification codes. Let $F_{im} = 1$ if patent document m contains the classification code i ($i = 1, \ldots, 30$) and $F_{im} = 0$ otherwise. The number of patents with classification code i is, therefore, given by $N_i = \Sigma_m F_{im}$. We can, thus, indicate the number of patents that are classified in both technological fields i and j as: $C_{ij} = \Sigma_m F_{im} F_{jm}$. By applying this count of joint occurrences to all possible pairs of classification codes, we obtain a square (30×30) symmetrical *matrix of co-occurrences* ($\boldsymbol{C}$), whose generic cell C_{ij} reports the number of patent documents classified in both technological fields i and j. This matrix of co-occurrences can then be used to derive a measure of relatedness between technological fields. Since we argued that no distinction ought to be made between primary and secondary classification, our measure of knowledge-relatedness has to be *symmetric* with respect to the direction linking technological fields. Moreover, such measure ought not to depend on the absolute size of the technological field (in terms of patent applications), otherwise we would over-estimate the knowledge links involving the larger fields.[16] A measure that fully meets these criteria is the *cosine index* S_{ij} (where i and j are two generic technological fields), which measures the angular separation between the vectors representing the co-occurrences of technological fields i and j with all the other fields:

$$S_{ij} = \frac{\sum_{k=1}^{30} C_{ik} C_{jk}}{\sqrt{\sum_{k=1}^{30} C_{ik}^2 \sqrt{\sum_{k=1}^{30} C_{jk}^2}}} \tag{5}$$

As the simple correlation coefficient, the cosine index provides a measure of the similarity between two technological fields in terms of their mutual relationships with all the other fields, with the advantage of being symmetric.[17] S_{ij} is the greater the more the two fields i and j co-occur with the same techno-

[14] See, for example, WIPO (1999).

[15] A further possible approach to the measurement of knowledge relatedness between technological fields is to use the information contained in the patent citations given in patent documents. In the official patent search reports the state-of-the-art related to the legal claims of each patent application is described using patent citations. By taking patent citations as units of analysis, the technological area of the cited patent is seen as the source from which technology spillovers originate, while the area of the citing patent is seen as the area absorbing the spillovers. However, investigations have shown that there are hardly any differences between the distribution of the cited and citing patents across technological areas. This is due to the fact that the classification of a patent document has the primary function to determine the areas of novelty for search purposes. Thus, the cited documents have almost always the same classification as the citing ones. Therefore, the use of patent citations does not provide additional information about knowledge spillovers or relatedness between areas of technology.

[16] It is easy to note that the observed value of co-occurrences C_{ij} is likely to increase with N_i and N_j, i.e. the number of patent applications in fields i and j, respectively. This implies that, if the size of the two technological fields is large (because N_i and N_j are large), one can expect to observe a fairly large number of patents co-classified in the two technological fields, even though the cognitive closeness among the two is very low. Conversely, if the size of the two technological fields is small (because N_i and N_j are small), one can expect to observe a relatively low number of co-occurences even though the two technologies are cognitively very close to each other.

[17] Of course, the cosine index is not the only measure of distance that could have been used. Among alternative methods, there is the absolute distance and the calculation of clusters. Regarding the former, the results one obtains are nearly the same when comparing relative indices (like the Revealed Comparative Advantage index), whereas they are most likely to be biased when (as in our case) comparing absolute patent numbers. Concerning the latter, cluster

logical fields. It is equal to one for pairs of technological fields with identical distribution of co-occurrences with all the other technological fields, while it goes to zero for pairs of technological fields that do not overlap.

Applying the cosine index calculation to a 30×30 matrix of the joint occurrences, we can generate a new matrix of the same size, whose elements are the various S_{ij} ($i = 1, \ldots, 30$; $j = 1, \ldots, 30$), which we call the *knowledge-relatedness matrix*. This matrix has been constructed using *all* patent applications to the EPO over the whole period 1982–1993, i.e. 721,260 observations (WORLD 1982–1993, Table 5).[18,19] The knowledge-relatedness matrix thus built will be used in the following section to test whether and to what extent firms extend their innovative activities across related technological fields. Here, we just point that most values are <20, with four outliers in between 20 and 30, four more between 30 and 50, and just one over 50. In addition, we notice that most of these outliers relate to technological classes related to the Chemical Sciences.

analysis allows to identify focal areas, but loose information about regular distributions. For these reasons, we chose to use the cosine index as an appropriate measure of distance.

[18] Searches and elaborations of data to calculate *the knowledge relatedness matrices*, as well as the methodology, were originally provided by the Fraunhofer Institute for Systems and Innovation Research (FhG-ISI—Karlsruhe), within the framework of the TSER-funded project "Innovation Systems and European Integration" (ISE), using the on-line version of the EPO database offered by the French database host Questel (EPAT). The delimitation of the time periods was done by the priority year as given in the patent application. The inventor country was used to distinguish between countries.

[19] In addition to that, since it was intended to examine whether knowledge relatedness remains stable or changes over time, and whether it is affected by differences among countries and the kind of technology produced there or whether it is relatively invariant across countries, we have also calculated six additional matrices. A first subset of three matrices is based on all patent applications to the EPO and refer each to one out of three subperiods (WORLD 1982–1985, WORLD 1986–1989 and WORLD 1990–1993). A second subset of three matrices refer instead to the three largest industrialised countries for the period 1990–1993 (US 1990–93, GERMANY 1990–1993 and JAPAN 1990–1993). No major difference emerges using world-wide patent data compared to individual country data. Furthermore, the relationships among technological fields remain highly stable over time. Therefore, in what follows, we will always make use of the matrix WORLD 1982–1993.

6. Knowledge-relatedness in the patterns of firms' technological diversification

In this section, we test the relative importance of knowledge-relatedness as an explanatory variable of individual firms' diversification choices, as compared to other patent-based indicators. To do so, we employ a logistic regression, where the dependent variable is a dichotomous one (presence versus absence in a certain technological class), and the key independent variable is our measure of knowledge-relatedness.[20]

Let ${}_hP_{i\neq j}$ be the probability that firm h, whose "core" technological field is j, is also active in technological field i (we define the "core" technological class below). Then a possible model of technological diversification pattern can be written as:

$$ {}_hP_{i\neq j} = f(CT_h, T_i, T_j) $$

where CT_h is a vector of technological characteristics of firm h; T_i a vector of characteristics of the target technology i; T_j a vector of characteristics of firm h's main technology j.

CT_h should accommodate for all of those variables that reflect one firm's propensity and/or opportunity to diversify its technological efforts, on the basis of some internal strategy or sheer size of its research effort (from which the chance to enjoy some research spillover). Two proxies could be derived from the EPO-CESPRI dataset, respectively,

SIZECLASS: number of technological classes in which firm h was reported to have at least one patent application, from 1982 to 1994; it ranges from 1 to 30.

SIZEPAT: overall number of patent applications held by firm h throughout the time interval 1982–1994; it ranges from 2 to over 9000, although most firms do not have >5 patents (see Section 2 above).

[20] It is quite important to point out that we are not testing any causal relationship going from knowledge relatedness to corporate technological diversification. It may be, in fact, that relatedness among technological fields is forged by firms' diversification activities. The purpose of our exercise is to assess to what extent firms' technological diversification tends to take place across related technological fields, whatever determines the degree of relatedness among these fields. We thank an anonymous referee for drawing our attention on this point.

Table 5
Knowledge-relatedness matrix based on co-occurrence of classification codes (WORLD 1982–1993) cosine indices × 100

	Technology classification fields[a]																													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1. Electrical engineering	100																													
2. Audiovisual technology	5.6	100																												
3. Telecommunications	6.0	15.4	100																											
4. Information technology	3.3	9.4	14.3	100																										
5. Semiconductors	11.6	3.8	3.7	5.4	100																									
6. Optics	6.8	9.8	6.7	3.8	10.2	100																								
7. Control technology	8.3	5.0	8.7	11.2	4.3	6.2	100																							
8. Medical technology	2.2	1.6	1.1	2.4	0.8	3.7	8.0	100																						
9. Organic chemistry	1.0	0.8	0.3	0.4	0.6	4.8	10.1	3.0	100																					
10. Polymers	5.7	1.9	0.4	0.3	2.4	9.5	1.8	5.3	14.8	100																				
11. Pharmaceutics	0.5	0.4	0.2	0.4	0.3	1.8	9.7	6.8	75.5	8.9	100																			
12. Biotechnology	1.0	0.7	0.9	1.4	0.6	1.8	25.7	4.2	47.7	5.9	47.9	100																		
13. Materials	10.1	1.6	0.5	0.4	7.9	4.0	2.3	2.1	4.3	8.6	2.6	1.9	100																	
14. Food chemistry	0.9	0.2	0.1	0.3	0.3	0.8	2.5	3.0	14.1	4.3	16.1	23.2	1.7	100																
15. Basic Materials chemistry	2.7	1.5	0.2	0.3	1.2	7.1	3.3	3.2	41.1	18.7	21.4	17.3	9.2	11.9	100															
16. Chemical engineering	3.1	0.8	0.6	1.3	1.8	2.2	9.5	6.1	15.2	8.3	7.3	8.2	16.5	8.1	17.1	100														
17. Surface technology	12.0	2.7	0.9	0.7	12.8	6.2	3.1	3.5	3.4	15.9	1.6	1.5	19.1	2.0	6.8	10.2	100													
18. Materials processing	4.5	2.1	0.4	0.7	1.8	5.3	3.0	5.0	3.6	26.8	1.8	2.1	9.7	2.3	11.3	9.8	18.9	100												
19. Thermal processes	7.8	0.6	0.8	0.7	1.9	1.0	6.0	2.0	0.9	1.2	0.5	1.2	12.4	1.7	3.3	10.4	3.7	3.6	100											
20. Environmental technology	1.9	0.3	0.2	0.3	1.0	1.1	2.6	2.5	4.5	4.8	2.1	3.8	15.1	3.4	9.7	37.3	6.7	4.8	13.3	100										
21. Machine tools	5.3	0.8	0.5	0.9	2.9	2.6	4.9	2.4	0.7	2.4	0.4	0.8	8.3	1.2	2.4	4.3	7.2	9.0	4.2	2.1	100									
22. Engines	4.7	0.5	0.8	0.7	0.9	0.6	5.2	1.9	0.4	0.5	0.3	0.8	2.8	0.3	1.1	4.2	1.9	1.3	8.4	5.8	2.8	100								
23. Mechanical elements	5.7	1.2	0.8	0.9	1.0	1.3	5.5	3.4	0.4	1.8	0.2	0.7	3.1	0.6	1.2	4.0	4.7	6.9	6.7	3.6	8.9	14.9	100							
24. Handling	3.5	2.7	1.5	3.4	2.1	5.6	6.0	3.6	1.2	3.8	0.7	1.1	2.1	3.7	2.3	6.6	10.2	11.6	2.7	2.2	7.3	1.5	6.7	100						
25. Food processing	1.2	0.5	0.3	0.6	0.4	0.6	3.6	2.1	2.1	1.5	2.2	5.0	1.2	14.4	3.8	6.2	1.8	2.6	2.6	3.1	3.5	1.2	3.4	5.7	100					
26. Transport	5.1	1.1	1.5	1.0	0.7	1.3	5.5	1.4	0.3	1.9	0.2	0.6	1.3	0.3	0.5	1.5	3.1	3.8	3.7	2.1	2.6	6.6	20.7	4.4	2.0	100				
27. Nuclear engineering	9.2	2.5	1.9	1.9	3.0	4.5	8.2	6.3	1.2	1.4	0.9	1.8	6.4	0.7	3.8	4.5	4.6	1.6	4.0	4.0	3.8	1.5	3.0	2.3	0.7	1.0	100			
28. Space technology	1.9	0.8	1.8	1.1	0.9	2.3	7.5	0.6	1.0	1.3	0.6	1.4	2.0	0.4	1.1	2.5	3.0	2.1	1.9	1.3	2.0	2.8	3.9	2.2	0.8	5.4	1.0	100		
29. Consumer goods	3.5	3.4	1.1	2.2	1.1	2.1	5.6	7.3	0.7	3.0	0.7	0.9	2.0	3.0	1.6	4.5	7.7	8.4	5.9	2.6	7.2	1.7	7.5	8.9	3.7	5.4	1.4	2.8	100	
30. Civil engineering	2.6	1.1	0.9	0.7	0.7	0.8	4.1	1.0	0.7	2.5	0.3	0.8	4.5	0.5	2.2	4.1	5.3	4.9	4.2	4.4	4.5	2.5	13.5	4.2	3.8	7.0	1.5	2.6	7.2	100

[a] Numerals 1–30 refer to the fields listed in the first column.

We expect both SIZEPAT and SIZECLASS to affect positively ${}_{h}P_{i \neq j}$. A cursory glance at the names of the largest innovators in the EPO-CESPRI dataset shows that the largest innovators are also the largest and eldest firms in each country. The role we give both to SIZEPAT and SIZECLASS is that of control variables (especially SIZECLASS). In particular, we expect knowledge-relatedness to affect less the larger (higher SIZEPAT) and more diversified (higher SIZECLASS) firms, due to the possibility that such firms will host a number of relatively independent business units. These provide alternative departure points for reaching a given technological class, rather than a single one, as we assume when positing the existence of just one "core" technological field for each firm (see below).

T_i contains all variables that may represent the opportunity to patent in target technology *i*, as well as the barriers to entry in such technology. Such barriers may derive from the existence of some indivisibility of research efforts at the firm level, or from the existence of national innovation system effects that favour firms from some countries. The variables we considered are:

- TOTPAT (_C, _W): total number of patents in class *i*, 1982–1994, in firm *h*'s country (TOTPAT_C; range of values is 26–12038) and in all the six countries (TOTPAT_W; range of values is 2485–36626);
- C4 (_C, _W): concentration index (share of patents in class *i* held by the largest four innovators), calculated both at the level of firm *h*'s country (C4_C) and for all the six countries (C4_W), 1982–1994; it varies between 0 and 1;
- VTRS: revealed technological advantage of firm *h*'s country in class *i*, 1982–1994. VTRS is a country specialisation index and it is calculated as the log of the share of world patents held by firm *h*'s country in class *i* over the same country's share of *all* world patents; it varies between −1 and 1.

Technological opportunity conditions are represented by TOTPAT (_C, _W). A positive sign is expected.

Barriers to entry and indivisibilities are represented by C4_C and C4_W. A negative sign of either C4_C or C4_W (which are highly collinear) is expected.

Finally, the role of national innovation systems is represented by VTRS. A positive sign of VTRS is expected: we suppose specialisation (VTRS > 0) to imply the existence of localised knowledge sources or institutional assets at the country level, which favour firms from that country that aim to access technology *i*.

T_j is exclusively composed by RELATEDNESS, our measure of knowledge-relatedness, between target class *i* and firm *h*'s core technological class *j*. The concept and measurement of knowledge-relatedness have been already discussed in Section 3; here, only normalised values from matrix world 1982–1993 have been used, so that RELATEDNESS ranges from 0 to 1. We expect RELATEDNESS to have a significant positive sign.

The definition of "core" technological class deserves some comments. We ought to find out in which class firm *h* first started innovating, and then map the firm's innovation efforts outside that class. However, EPO patents have to be regarded as left-censored innovation proxies, since none of them was granted before 1978, well after most firms in our dataset had started innovating. In addition, it may well be that some firms which have started innovating in some technological area then focus on a new one, and depart from there for their further diversification efforts. So, any "chronological" definition of core technological class is empirically unfeasible and, at the same time, not necessarily sound from a theoretical viewpoint. Therefore, we have identified as firm *h*'s core technological class the one in which the firm is relatively more specialised in its technological activities. However, this cannot be identified as the one that records the highest number of firm *h*'s patents, because patents represent an appropriability instrument whose effectiveness differs widely across technologies. Innovation in some technologies is better protected by other appropriability means (e.g. secrecy or time-to-market), so that similar innovation efforts in two different technologies may result in a widely different number of patents in the corresponding technological classes. By comparing the number of patents one could end up electing as a firm's "core" technological class simply the one with the higher "patent:innovation" ratio, for the same amounts of innovative activity. A better criterion is that of comparing firm *h*'s *shares* of patents in the various technological classes (i.e. the number of patents held by the firm in each class, over total number of patents at the world level in that class). We have then chosen the class with the highest share as the firm's "core" one.[21]

Due to our a priori on the role of SIZECLASS (see above), we expected relatedness to interact negatively with it, or to reduce its impact on ${}_hP_{i\neq j}$, if tested on subsamples of firms with high values of SIZECLASS.

We run two different sets of regressions. In the first set each firm appears 29 times, i.e. as many times as many technological classes exist which are different from its core one (i.e. the dependent variable is ${}_hP_{i\neq j}$, with $i = 1, \ldots, 30$). A basic regression (with alternative model specifications) has been run, which was based on the whole sample of firms (as described in Tables 2 and 3 above), as well as a number of additional ones, each of them referred to a different subsample, based upon the values of SIZECLASS. Table 6 reports the results for the basic regression (all firms), and for four of the additional ones, namely those that refer to firms diversified in 2–5 classes (SIZECLASS[2,5]), 6–10 classes (SIZECLASS[6,10]), 11–20 classes (SIZECLASS[11,20]), and 21–30 classes (SIZECLASS[21,30])[22]. In particular, Table 6 refers to a model specification which employs TOTPAT_W and C4_C rather than TOTPAT_C and C4_W (no relevant differences emerged when adding the latter, or using them as substitutes to the former). SIZEPAT and TOTPAT_W appear in logarithmic form to allow for non-linearity in their relationship with ${}_hP_{i\neq j}$.

When testing for the existence of interaction effects we discovered a great number of them to be significant. The stepwise procedure we employed for their identification always pointed at the interaction between SIZECLASS and RELATEDNESS as a key one, but also included different interaction effects depending on the chosen subsamples. The number and variability of such interaction effects pose interpretation problems beyond our current reach, so Table 6 reports (along with the Main Effect models) only the models with SIZECLASS-RELATEDNESS interaction effects. The profile of ${}_hP_{i\neq j}$, as a function of RELATEDNESS resulting from these limited interaction models, however, resembles very closely the profile obtained from both the full interaction ones and the model specification that includes only the SIZECLASS–RELATEDNESS interaction effect, but does not resemble the profile resulting from any model which allow for the main effects only (see comments to Fig. 3 below).[23]

In the Main Effect model in Table 6 all the signs of the independent variables are as expected. The role of knowledge-relatedness is positive and highly significant. In addition, the positive sign of VTRS indicates that firms tend to diversify in those technologies in which a country has advantages. On the other hand, firms face difficulties to enter in those classes in which technological activities are already highly concentrated and therefore barriers to entry of various types are high (negative sign of C4_C).

In addition, the main effect of relatedness increases when moving from the SIZECLASS[2,5] subsample up to the SIZECLASS[21,30] one. This suggest that the influence of relatedness on the probability for a firm to diversify in a given field *increases* with the extent of the firm's diversification. That is, some knowledge-related coherence appears in firms' diversification patterns, which does not fade away with the extent of diversification. In particular, it is possible that firms which diversify in two or a few more technological classes, although reaching close classes, do not necessarily reach the closest one. At the same time, firms which diversify in a higher number of classes do not leave any "gap" in the relatedness hierarchy that separates their core technological class

[21] However, even this measure is not immune from drawbacks. In particular, the share of patents in a given class may be inversely related to the number of firms and/or patents in the small class. Indivisibilities exist, such that niche technologies can host a relatively small number of firms and patents, so that innovators have high shares, even if their engagement in those technologies represent just a minimal part of their whole innovation effort. Our 30-class classification scheme, however, does not contain very small classes, with the possible exceptions of classes 27 (nuclear technology) and 28 (space technology). Notice that similar remarks may apply also to other share-based variables, such as C4 and VTR.

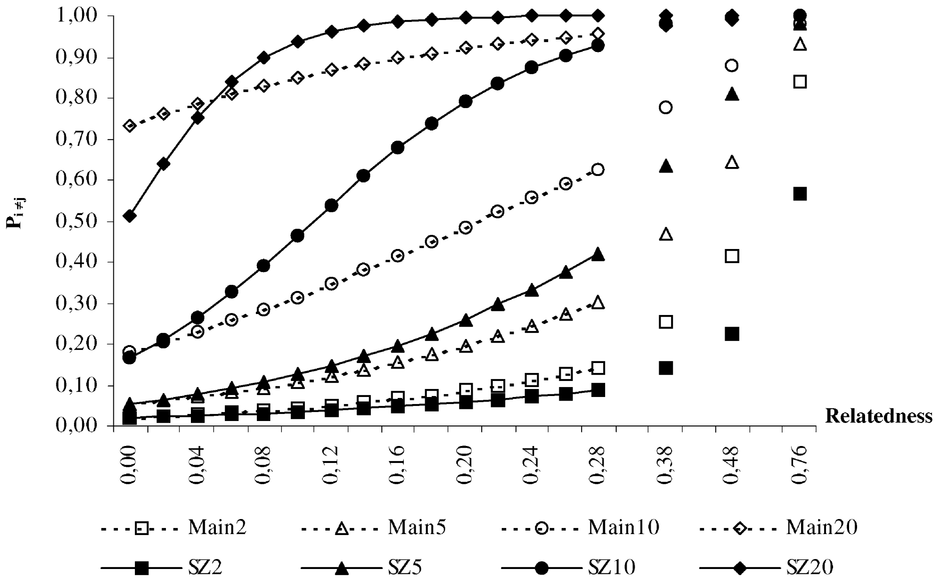
[22] Other regressions refer to subsamples for SIZECLASS equal to 2, 3, 4 and 5, respectively, as well fro each of the siz countries in the EPO-CESPRI database (all SIZECLASS). All of them give results which are very similar to those reported in Table 6.

[23] In particular, interpretation is made easier by the absence of changes in the sign of the main effect parameters (in particular RELATEDNESS), due to the insertion of too many interaction effects. Notice that such changes did not affect the overall profile of ${}_hP_i \neq_j$, as a function of RELATEDNESS, but simply made it more difficult to see the relationship between the two, without plotting it on a complex graph like the one in Fig. 3.

Table 6
Technological diversification as a function of knowledge-relatedness logistic regression/selected model specifications (six countries, 1982–1993)[a]

	All firms		SIZECLASS[2,5]		SIZECLASS[6,10]		SIZECLASS[11,20]		SIZECLASS[21,30]	
	Main effects	Interaction effects	Main effects	Interaction effects	Main effects	Interaction effects	Main effects	Interaction effects	Main effects	Interaction effects
INTERCEPT	−6.48	−6.90	−7.35	−8.11	−5.61	−5.41	−5.40	−4.94	−10.19	−6.17
RELATEDNESS	7.19	3.10	6.47	2.91	11.00	7.13	17.74	6.06[b]	33.31	–
SIZECLASS	0.22	0.33	0.56	0.80	0.23	0.20	0.18	0.16	0.34	0.17
VTRS	1.19	1.35	1.30	1.30	1.06	1.06	0.97	0.96	0.83	5.81[b]
C4_C	−2.59	−2.98	−2.90	−4.15	−1.92	−1.91	−1.87	−1.84	−1.50	−1.47
LogTOTPAT_W	1.11	1.23	1.20	1.52	0.99	0.99	0.92	0.92	0.99	–
LogSIZEPAT	0.19	0.24	0.01 NS[c]	0.22	0.00 NS	–	0.03 NS	–	0.12[b]	–
D1 (SIZECLASS× RELATEDNESS)		1.17		1.28		0.52		0.85		1.45
D6 (SIZECLASS× VTRS)		−0.03		–		–		–		−0.21[b]
D7 (SIZECLASS× C4_C)		0.08		0.40		–		–		–
D8 (SIZECLASS× LogTOTPAT_W)		−0.02		−0.10		–		–		0.04
D9 (SIZECLASS× LogSIZEPAT)		−0.02		−0.07		–		–		0.01
−2LogL	177800.44	175113.97	122053.82	121702.98	33360.96	33353.33	16280.46	16265.24	2586.32	2576.00
−2LogL (intercept only)	244554.77	244554.77	145852.74	145852.74	39569.05	39569.05	20226.79	20226.79	3483.95	3483.95
Predicted probability/concordant (%)	85.0	85.8%	80.6	80.7	77.4	77.4	79.0	79.1	84.3	84.4
No. of observations (no. of firms × 29)	390514	390514	334892	334892	37410	37410	14645	14645	3567	3567
No. of logit = 1 (% of total observations)	36954 (9.5)	36954 (9.5)	18985 (5.7)	18985 (5.7)	8288 (22.2)	8288 (22.2)	6797 (46.4)	6797 (46.4)	2884 (80.9)	2884 (80.9)

[a] All coefficients significant at 99%. Source: EPO–CESPRI database.
[b] Significant at the 95%.
[c] NS: non significant.



Notes

1. Main*X* refer to AllFirms/Main Effects model of Table 6, with:
   SIZECLASS=*X*
   VTRS=Avg. Value (All Firms)
   C4_C=Avg.Value (All Firms)
   LogTOTPAT_W=Avg.Value (All Firms)
   LogSIZEPAT=Avg.Value (Firms with SIZECLAS=*X*)
2. SZ*X* refer to AllFirms/ Interaction effects model of Table 6, with the other variables as in note 1.

Source: EPO-CESPRI database.

Fig. 3. Firm h's probability to diversify into class $i \neq j$ as a function of knowledge-relatedness between i and j (j = firm h's core class).

from the furthest they reach. These firms do not jump from their core class to the farther ones: rather they fill in all the classes in between.

Signs of the independent variables do not change when moving to the Interaction Effect model. The interaction of relatedness with SIZECLASS is a positive one.

These results are confirmed by Fig. 3, which plots the probability of patenting outside one's core technological field (${}_hP_{i\neq j}$,) as a function of knowledge-relatedness. Two sets of four curves are plotted: the white-dotted curves are based on the estimated parameters of the Main Effect model in Table 6, while the black-dotted curves illustrate the Interaction Effect model. The horizontal axis reports the values of relatedness from 0 to 0.28 in a continuous fashion, and then three of the few extreme values pointed out in Section 5. Each curve within a group is based on a different value of SIZECLAS (2, 5, 10 and 20, respectively), while all curves are based on average values of the other control variables. For example, the curve labelled Main2 represents the probability to patent in any class i other than one's core class j, for all firms active in no more than two fields, on the basis of the estimates of the Main Effect model: as expected the probability to be active in i increases with the relatedness of i–j. If we compare Main2 with Main5, the latter lies above the former for all values of knowledge-relatedness. Both curves tend to one as knowledge-relatedness tends to one.

More interestingly, we notice that adding the SIZECLASS–RELATEDNESS interaction effects to the main ones change significantly the profile of ${}_hP_{i\neq j}$, compared to the Main Effect model. In particular, firms with SIZECLASS over 5 appear to increase their chances to diversify more than the other ones. Similar plots, which report the results of regressions carried on the subsamples described in Table 6, suggest similar conclusions.

The latter are also confirmed by the second set of logistic regressions. Rather than pooling all values of ${}_hP_{i\neq j}$, for each firm, in this second set separate regressions for each technological class are run, for a total of 30. In all cases RELATEDNESS and SIZECLAS are significant and interact positively, (although, inevitably, C4_C rather than TOTPAT_C[24] are often non significant). We do not report the results in full for sake of brevity (they are available on request).

6.1. Possible extensions and applications

The results reported above provide support to the notion that firms are coherent in terms of the technologies they are active in, and identifies knowledge as a key factor at the base of this coherence. The analysis carried out in this paper allows several potential applications and extensions, both at the theoretical and at the empirical level.

At the theoretical level, our findings point to the need for more research aimed at disentangling the possible determinants and dimensions of 'relatedness'. As argued above, relatedness may be the result of quite different factors: proximity, commonality or complementarity. A finer grained analysis should therefore be directed to clarify the relationship between the possible sources of knowledge-relatedness, on one hand, and the properties of the knowledge base and the characteristics of firms' product lines, on the other. In addition, more theoretical (and empirical) efforts should be devoted to exploring the extent to which relatedness derives from intrinsic properties of the knowledge base and products underpinning firms' innovative activities, as opposed to the possibility that relatedness is purposively shaped by firms attempting to broaden the scope of their activities. Relatedly, research aimed to model firms' diversification choices would also help to disentangle the *unintended* versus the *intended* learning processes that are at the base of diversification, and the relationship between knowledge-relatedness, sunk costs and switching costs of moving away from established technologies.

On the empirical side, further research is needed to validate and extend our analysis. In the first place, other measures of knowledge-relatedness should be tested and compared to the one used here, i.e. co-classification codes of patent documents. Among them, one can think of patent citations, non-patent literature citations (scientific publications), and teams of co-inventors. In the second place, other indices and metrics should be also adopted in order to assess the robustness of our results. At a more substantial level, we believe that the basic methodology developed in this paper could be fruitfully applied to shed light on a rather important issue, namely, the *technological trajectories* followed by firms over time. Two contrasting hypotheses have been put forward in the recent literature. One suggests that persistently innovative firms increasingly *focus* their research on an increasingly smaller number of technologies, thus narrowing down their patterns of specialisation. By contrast, other authors claim that (large) firms are constantly engaged in opening up windows on new technologies, or entering into technologies that are complementary, or in closely related in terms of knowledge, or subject to spillover effects. According to this second hypothesis, a *widening* process is at work, with firms broadening the number of technologies they need or wish to master (Granstrand, 1998; Granstrand et al., 1997). Although it goes beyond the objectives of this paper to enter into this debate, we believe that mapping the relationships and measuring the extent of technological relatedness, along the ways proposed in this paper, would greatly contribute to discriminating among competing hypotheses (see also, Breschi and Malerba, 2000; Breschi et al., 2002).

7. Conclusions

The paper has proposed that relatedness in knowledge is a key factor in affecting firms' technological diversification. First, in a descriptive fashion it has examined the extent of technological diversification in six major advanced countries. It has found that

[24] Notice that TOTPAT_W and C4_W could not be used in this second exercise, their value not changing within the same technological class.

technological diversification is quite widespread among small as well as major innovators. Second, the paper has tested the hypothesis that relatedness in technologies is a major driver of firms' technological diversification (knowledge-relatedness hypothesis) versus the hypothesis that technological diversification occurs randomly (random hypothesis). It has found evidence in favour of the relatedness hypothesis: firms extend their innovative activities across knowledge-related technological fields as a consequence of learning processes (either unintended in terms of spillovers or intended in terms of local learning) and of knowledge features and links (due to scope, complementarity or the generic nature of knowledge). Even the most technologically diversified firms patent in closely related fields. A particularly interesting result is that larger diversifiers are more "coherent" in terms of knowledge-relatedness of their technological activities than smaller diversifiers. This could be due to the fact that, as firms expand their activities, they gradually increase their knowledge coherence by filling systematically all the classes that are related to those classes they started with, thus filling all the possible "technological gaps" between classes. We believe that our results pave the way for further research, both in terms of methodological perspectives (how to measure knowledge and productive relatedness?), and for a better understanding of firms' technological strategies.

Acknowledgements

This paper draws upon the research project *Innovation Systems and European Integration* (ISE), funded by the Targeted Socio-Economic Research (TSER) programme of the European Commission (DGXII) under the 4th Framework Program. (contract no. SOE1-CT95-1004). Useful comments and support from all the participants to the ISE project are gratefully acknowledged.

References

Antonelli, C., 1995. The Economics of Localised Technological Change and Industrial Dynamics, Kluwer Academic Publishers, Dordrecht.

Arora, A., Gambardella, A., 1994. The changing technology of technological change: general and abstract knowledge and the division of innovative labour. Research Policy 23, 523–532.

Atkinson, A.B., Stiglitz, J.E., 1969. A new view of technical change. Economic Journal 79, 573–578.

Breschi, S., Malerba, Fr., 2000. Diversification and Specialisation in Innovative Activities. An Analysis of Patenting Activities of Electronic Firms, WP TSER Dynacom, Scuola Sup. S. Anna, Pisa.

Breschi, S., Lissoni, F., Malerba, F., 2002. The empirical assessment of firms' technological coherence: data and methodology. In: Cantwell, J., Gambardella, A., Granstrand, O. (Eds.), The Economics and Management of Technological Diversification, Routledge.

Bresnahan, T.F., Trajtenberg, M., 1995. General purpose technologies 'Engines of growth'? Journal of Econometrics 65, 83–108.

Cantwell, J., Andersen, B., 1996. A statistical analysis of corporate technological leadership historically. Economics of Innovation and New Technologies 4, 211–234.

Caves, R.E., 1981. Diversification and seller concentration: evidence from changes. The Review of Economics and Statistics 63, 289–293.

Cefis, E., 1996. Is there any persistency in innovative activities? Working Paper, Department of Economics, University of Trento, Trento.

David, P.A., 1975. Technical Choice, Innovation and Economic Growth, Cambridge University Press, Cambridge, UK.

Dosi, G., 1997. Opportunities, incentives and the collective patterns of technological change. Economic Journal 107, 1530–1547.

Engelsman, E.C., van Raan, A.F.J., 1991. Mapping of Technology, A First Exploration of Knowledge Diffusion Amongst Fields of Technology, Policy Studies on Technology and Economy (BTE) Series, No. 15. The Hague.

Engelsman, E.C., van Raan, A.F.J., 1992. A patent-based cartography of technology. Research Policy 23, 1–26.

Geroski, P., Van Reenen, J., Walters, C.F., 1997. How persistently do firms innovate? Research Policy 26, 33–48.

Granstrand, O., 1998. Towards a theory of the technology-based firm. Research Policy 27 (5), 467–491.

Granstrand, O., Patel, P., Pavitt, K., 1997. Multi-technology corporations: why they have distributed rather than distinctive core competencies. California Management Review 39 (4), 8–25.

Griliches, Z., 1979. Issues in assessing the contribution of research and development to productivity growth. Bell Journal Economics 10 (1), 92–116.

Griliches, Z., 1991. Patent statistics as economic indicators: a survey. Journal Economic Literature 28, 1661–1707.

Griliches, Z., 1995. R&D and productivity: econometric results and measurement issues. In: Paul Stoneman (Ed.), Handbook of the Economics of Innovation and Technological Change. Blackwell, Oxford.

Grupp, H., 1996. Spillover effects and the science base of innovations reconsidered: an empirical approach. Journal Evolutionary Economics 6, 175–197.

Henderson, R., Cockburn, I., 1996. Scale, Scope, and Spillovers: the determinants of research productivity in drug discovery. Rand Journal Economics 27, 32–59.

Hinze, S., Reiss, T., Schmoch, U., 1997. Statistical Analysis on the Distance Between Fields of Technology, FhG-ISI (Karlsruhe): ISE Working Paper.

Jaffe, A.B., 1986. Technological opportunity and spillovers of R&D: evidence from firms' patents, profits, and market value. American Economic Review 76 (5), 984–1001.

Jaffe, A.B., 1989. Characterising the technological position of firms, with application to quantifying technological opportunity and research spillovers. Research Policy 18, 87–97.

Lemelin, A., 1982. Relatedness in the patterns of interindustry diversification. The Review of Economics and Statistics 64, 646–657.

MacDonald, J.M., 1985. R&D and the directions of diversification. The Review of Economics and Statistics 67, 583–590.

Malerba, F., 1992. Learning by firms and incremental technical change. Economic Journal 102, 845–859.

Malerba, F., Orsenigo, L., Peretto, P., 1997. Persistence of innovative activities, sectoral patterns of innovation and international technological specialisation. International Journal Industrial Organisation 15, 801–826.

Milgrom, P., Roberts, J., 1990. Modern manufacturing: technology, strategy and organization. American Economic Review 80 (3), 511–528.

Nelson, R.R., Winter, S.G., 1982. An Evolutionary Theory of Economic Change, Harvard University Press, Cambridge, MA.

Patel, P., Pavitt, K., 1995a. Technological Competencies in the World's Largest Firms: Characteristics, Constraints and Scope for Managerial Choice, IIAASA WP-95-66, Laxenburg.

Pavitt, K., 1998. Technologies, products and organisation in the innovating firm: what Adam Smith tells us and Joseph Schumpeter does not. Industrial and Corporate Change 7 (3), 433–452.

Penrose, E., 1959. The Theory of the Growth of the Firm, Basil Blackwell, Oxford.

Putnam, J., Evenson, R. E., 1994. Inter-sectoral Technology Flows: Estimates from a Patent Concordance with an Application to Italy, Mimeo, Yale University.

Scherer, F.M., 1982. Interindustry technology flows in the United States. Research Policy 11, 227–245.

Scott, J.T., 1993. Purposive Diversification and Economic Performance, Cambridge University Press, New York.

Teece, D.J., 1982. Toward an economic theory of the multiproduct firm. Journal of Economic Behaviour Organization 3, 39–63.

Teece, D.J., Rumelt, R., Dosi, G., Winter, S.G., 1994. Understanding corporate coherence: theory and evidence. Journal of Economic Behaviour Organisation 23, 1–30.

Verspagen, B., 1997. Measuring inter-sectoral technology spillovers: Estimates from the European and US patent office databases. Economic Systems Research 9 (1), 49–67.

WIPO, 1994. International Patent Classification: Guide, Survey of Classes and Summary of Main Groups, 6th Edition, Vol. 9. World Intellectual Property Organization, Geneva.

WIPO, 1999. IPC Reform Project File. Elaboration of Multiple Classification in the IPC and Investigation of Possible Bearing of Such Rules on General Principles of Classifying Disclosed in the Guide to IPC, Document IPC/R 4/99 from: http/www.wipo.int/eng/main.htm, World Intellectual Property Organization, Geneva.