NBER WORKING PAPER SERIES

THE EFFECTS OF MARKET DEMAND, TECHNOLOGICAL OPPORTUNITY AND RESEARCH SPILLOVERS ON R&D INTENSITY AND PRODUCTIVITY GROWTH

Adam B. Jaffe

Working Paper No. 1432

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 1984

The research reported here is part of the NBER's research program in Productivity and project in Productivity (World Economy). Any opinions expressed are those of the author and not those of the National Bureau of Economic Research.

NBER Working Paper #1432
August 1984

The Effects of Market Demand, Technological Opportunity and Research Spillovers on R&D Intensity and Productivity Growth

ABSTRACT

This paper uses sales and patent distribution data to establish the market and technological "positions" of firms. A notion of technological proximity of firms is developed in order to quantify potential R&D spillovers. The importance of the position variables and the potential spill-over pool in explaining R&D intensity, patent productivity and TFP growth is explored.

I find that both technological and market positions are significant in explaining R&D intensity, and that the technological effects are significant in explaining patent productivity. I cannot distinguish between the two effects in explaining TFP growth. Spillovers are important in all three contexts. Firms in an area where there is a high level of research by other firms do more R&D themselves, they produce more patents per R&D dollar, and their productivity grows faster, even controlling for the increased R&D and patents. These effects are present controlling for both industry and technological position effects.

Adam B. Jaffe
National Bureau of Economic Research
1050 Massachusetts Avenue
Cambridge, Ma 02138

(617) 868-3900

I. Introduction[1]

This paper uses company cross-section data to investigate two interrelated questions regarding the determinants of industrial R&D activity and productivity growth. The questions are: (1) what is the relative importance of market conditions and exogenous technological conditions in determining the allocation of resources to research, and in explaining the growth of productivity; and (2) are there observable "spillover effects" among firms' research programs.

There is a long history of debate in economics on whether innovation is more often the result of the "pull" of market forces or of the "push" of exogenous technological factors. Schmookler (1966) argued that, looking across industries over time, the effect of expected market size on the expected return was the primary determinant of the allocation of resources to invention.[2] Rosenberg (1974 and 1983), while not denying the importance of the factors cited by Schmookler, argues that factors on the cost or supply side of invention are also of major importance in explaining both the allocation of resources to invention and the resulting growth of productivity. These supply-side factors include variations in both the intrinsic difficulty of innovation in different technical areas and also variations in the state of general knowledge in different areas. Both of these factors are often summed up in the phrase "technological opportunity." Scherer (1965 and

---

1. I am indebted to Profs. Zvi Griliches and Richard Caves for their ongoing advice and counsel, and to Prof. Mark Schankerman, Jim Hines, Sumanth Addanki and John Bound for helpful suggestions. Financial support was provided by the National Science Foundation through grant PRA81-08635 and by the Sloan Foundation.

2. Myers and Marquis (1969) found that most successful innovations were perceived by the firms involved as having been market driven. This and other related studies are critiqued in Rosenberg (1983).

especially 1982a) finds that dummies for broad technological classes explain much of the variance in innovative intensity across industries and argues that this is evidence for the importance of technological opportunity. In this paper, I use independently derived data on firms' "market positions" and "technological positions" to distinguish demand-pull and technological opportunity effects.

The theoretical possibility and potential importance of R&D spillovers are more recent concerns in the literature. They have been discussed by Griliches (1979), Reinganum (1981), and Spence (1984).[3] Griliches points out that the extent of spillovers is crucial to assessing the contribution of R&D to productivity growth, and Spence emphasizes that in R&D-intensive industries the extent of spillovers is an element of market structure with important implications for conduct and performance. Despite this recognition of the importance of spillovers, there is little empirical evidence on their actual significance.

Like technological opportunity, spillovers affect the cost of innovation. If appropriability is imperfect, then having lots of other firms doing research in the same areas as you may reduce the cost of innovating for you. This positive externality should be observable as a correlation between, on the one hand, the firm's innovative output and productivity growth, and, on the other hand, the

---

3 . The term "spillovers" is used by different writers to describe slightly different phenomena. I use it to describe a transfer of knowledge from one firm to another unmediated by any market transaction. I thereby exclude the phenomenon whereby improved products are sold at prices that do not fully reflect their higher quality, thereby "spilling" research benefits from suppliers to customers. See Scherer (1982b) and Griliches and Lichtenberg(1984a).

intensity of research in its "neighborhood" (to be made precise below). This correlation should remain even after controlling for the firm's own R&D.

This positive externality is technological: it derives from the inherent public good aspect of knowledge. There are also reasons to expect negative externalities among research programs, generated by competition. Because of patent laws, other firms' research may reduce your innovative output, particularly if innovative output is measured by patents. Similarly, most firms face downward sloping demand curves, which shift inward when a competitor lowers price or improves quality. This generates a negative effect of other's research on the firm's productivity, if productivity is measured in terms of revenues. Empirically, it would be difficult to distinguish separately the positive and negative effects. In this paper I simply estimate partial correlations that capture the net effect of the two.

It should be emphasized that this paper examines the effect of variations in spillovers, not variations in the conditions of appropriability. The distinction can be clarified by reference to the model of Spence(1984). Assume that for every dollar a firm spends on R&D, a fraction θ "spills over" and is freely available to any firm. The appropriability environment is then parameterized by θ. Low θ represents easy appropriation; θ approaching unity represents R&D as a pure public good. Increasing θ (reducing the extent of appropriation) has two effects: the private rate of return to R&D falls because of

the inability to capture the returns, but the industry-wide productivity of R&D rises because of the fuller exploitation of the public good nature of knowledge.

Spence's firms are identical, so they all enjoy spillovers to the same extent; in this framework, Spence examines the effect of changing θ. In this paper, I assume, in effect, that θ is the same for all firms, but that firms differ in their R&D intensity and their technological areas of interest. I examine the effect of the variations in the "pool" of spilled research available to different firms, these variations being generated by variations in the intensity of other firms' research in the relevant technological areas.

To measure this pool of relevant spilled research (and to test for technological opportunity effects), we need a framework for characterizing and differentiating the research interests of firms. The next section of the paper establishes such a framework. The following three sections present empirical results for R&D intensity, patent production and productivity growth. Concluding observations follow.

II. Technology Space, Market Space, and Technological Proximity

The typical large manufacturing firm is somewhat diversified both in terms of the products it sells and the research it pursues. These two patterns of diversification are not unrelated, but they are distinct. Even the seller of a single product like an automobile will typically do research in diverse areas, such as aerodynamics, engines,

and structural properties of materials. For the purpose of this paper, I do not try to understand these patterns of diversification, or their mutual relationship. I seek only to characterize the firm's research interests and market position in a way that takes them into account.

To do this, I construct two conceptually analagous distribution vectors for each firm. Firm i's technological position vector $f_i=(f_{i1}...f_{iK})$ indicates the fraction of the firm's research effort devoted to the K diverse technological areas. It's market position vector $g_i=(g_{i1}...g_{iL})$ indicates the fraction of its sales that go to the L distinct markets.[4]

For concreteness, we can visualize these vectors as locating the firm in a K-dimensional technology space and an L-dimensional market space. The question of the importance of technological opportunity then becomes: Does position in technology space matter in explaining patterns in the allocation of resources to research and in productivity growth? The importance of demand-side effects can be explored by asking the same question for position in market space.

In this framework, the existence of R&D spillovers implies that a firm's performance is affected by the research activity of its neighbors in technology space. To make this notion operational requires significant additional structure. I assume that the total relevant activity of other firms can be summarized by a "potential spillover pool" that is simply a weighted sum of other firms' R&D,

---

4. The characterization of technological areas and markets used to make this approach operational are discussed below.

with weights proportional to the proximity of the firms in technology space. To measure the proximity of firms i and j, I use the angular separation or uncentered correlation of the vectors f_i and f_j:

$$P_{ij} = \frac{f_i f_j'}{[(f_i f_i')(f_j f_j')]^{1/2}} \qquad (1)$$

Thus, in constructing the pool for firm i, firms whose position vector is identical to f_i get a weight of unity; firms whose vector is orthogonal get zero weight, and others get an intermediate weight that depends on the degree of overlap of the two firms' research interests.[5]

In this formulation, firms with rather diverse research interests potentially benefit to a small extent from the research of virtually all other firms. (That is, there may be no j for which P_{ij} is zero.) This may seem implausible. An alternative view is that firms are aware of the activities of a fairly small number of technologically similar firms; developments by others, even though potentially relevant, are simply not likely to be noticed. To explore this possibility, the firms were clustered into groups based on their technological positions, and the total pool described above was partitioned into the part coming from other members of the same cluster, and the part coming from outside the cluster. In the empirical work below, the differential effects of these two parts is explored.

Before going on to describe the empirical results, I pause to discuss the data used to construct the position vectors and the spillover pool. The technological position is based on the distribu-

---

5. Bernstein and Nadiri (1983) measure spillovers using a cost function approach and time series data for the chemical industry. Their pool variable is the unweighted sum of the R&D of all other firms in the industry.

tion of the firm's patents over the 328 technology classes used by the U.S. Patent Office. This information is available for the patents granted between 1969 and 1981 to about 1700 manufacturing firms in the R&D panel that has been assembled recently at the NBER. This dataset is a marriage of Compustat and Patent Office data that is documented in Cummins, et al (1984) and Bound, et al (1984). The companies in the dataset were granted about 260,000 patents over the period. The average firm has one or more patents in about 20 of the 328 classes. The classes themselves vary greatly in importance, from "Chemistry, carbon compounds" with 20,000 patents taken by 340 different firms to "Bee culture" which has one patent. (There are eight classes in which these firms took no patents, including "Land vehicles, animal draft" and "Whips and whip apparatus.") To make the distribution vectors empirically usable, the 328 classes were grouped into 49 categories. This grouping was essentially ad hoc, based on the names, with more aggregation of classes with few patents, and less aggregation of those that had many.

For this paper, a subsample of 573 firms was selected. These firms all reported R&D in 1976, received at least 10 patents over the period, and are also available on the Harvard Business School PICA database. The first two requirements bias the sample toward technology intensive firms, and the latter toward large firms. The firms had average 1976 sales of $1.4 billion, they average 29.5 patents per year, and they reported average R&D of $25.1 million in 1976. The

TABLE 1

PATENT CATEGORIES WITH TOTALS FOR CLUSTERING SAMPLE, 1965-72

CATEGORY	NUMBER OF PATENTS	NUMBER OF COMPANIES WITH ANY PATENTS	NUMBER OF COMPANIES HAVING THIS AS PRIMARY CATEGORY
ACOUSTICS	424	86	2
ADHESIVES AND COATINGS	5093	340	21
AERONAUTICS	417	61	3
AMMUNITION & EXPLOSIVES	842	93	4
APPAREL	645	130	4
BOATS, SHIPS & AQUATIC DEV.	269	67	1
CHEMISTRY, ANAL. & PHYS.	1404	209	4
CHEMISTRY, CARBON	10408	220	45
CHEMISTRY, ELECTROCHEMISTRY	5779	245	10
CHEMISTRY, HYDROCARBONS	2666	110	6
CHEMISTRY, INORGANIC	2149	140	3
CHEMISTRY, ORGANIC	5900	146	15
CLEANING AND ABRADING	1252	216	11
COMBUSTION	1056	147	4
COMPOSITIONS	3679	208	11
CUTTING	991	212	11
DRUGS	2442	132	6
ELEC COMPUTERS & DATA PROC.	4466	198	12
ELEC TRANSMISSION & SYSTEMS	7609	291	37
ELEC MOTORS & GENERATORS	2706	205	6
ELEC COMMUNICATION	5629	230	20
FARMING	229	77	0
FLUID HANDLING	6866	396	45
FOOD	1734	152	28
MACHINE ELEMENTS	2447	238	3
MEASURING,TESTING & SIGNALING	3987	323	16
MEDICAL	1724	196	10
METALS & METAL WORKING	5703	334	28
MISC ARTICLE HANDLING	5760	387	32
MISC ARTICLES	2998	249	9
MISC CONSUMER GOODS	1410	194	16
MISC ELECTRONIC DEVICES	2458	128	1
MISC HARDWARE	1376	230	7
MISC MACHINERY	855	134	7
NUCLEAR	213	21	0
OPTICS	3744	179	7
PAPER MAN & FIBER PROC.	571	125	8
PIPES & JOINTS	1214	226	2
POWER PLANTS (NON-ELECTRIC)	4145	311	20
PRINTING & TYPEWRITING	954	139	6
RADIANT ENERGY	5254	244	13

TABLE 1 (Continued)

PATENT CATEGORIES WITH TOTALS FOR CLUSTERING SAMPLE, 1965-72

CATEGORY	NUMBER OF PATENTS	NUMBER OF COMPANIES WITH ANY PATENTS	NUMBER OF COMPANIES HAVING THIS AS PRIMARY CATEGORY
RECEPTACLES & PACKAGES	2190	286	17
REFRIDGERATION & HEAT EXCH.	3260	264	15
SERVICES	197	61	1
STATIC STRUCTURES	1282	269	13
SYNTHETIC RESINS	4382	152	6
TEXTILE MAN & TREATMENT	1699	146	7
VEHICLES	2948	198	13
WELLS & EARTH WORKING	2230	129	7

technological position is based on all patents applied for by 1972.[6] Table One lists the 49 patent categories with some statistics for the sample used in this paper.

The market position vectors were constructed using the firm's distribution of shipments over 4-digit manufacturing SIC's in 1972. These data were taken from the PICA database at Harvard Business School, described in Shesko (1982). They are based on the assignment of the firms' establishments to SIC's, and the aggregation of establishment shipments. I aggregated these 450 SIC's up to the 19 industry groups used in Bound, et al, which were designed to correspond to the product field categories used by the NSF in its R&D surveys. Table 2 shows these industries with some statistics; the definition of the industries by constituent SIC's is given in Appendix A.

To give some idea of what these vectors look like, Figure One shows the distribution of the Herfindahl indices (the sums of the squared fractions) of the patent category and industry shares for each firm. It shows that, in this data, firms are far more diversified in technology space than in market space. The median patent class Herfindahl is about .18, while for the industry fraction vectors it is about .68. This is not due primarily to the larger number of patent categories; calculating the industry Herfindahl at the 3-digit level

---

6. Note that only patents that were ultimately granted are included in the database.

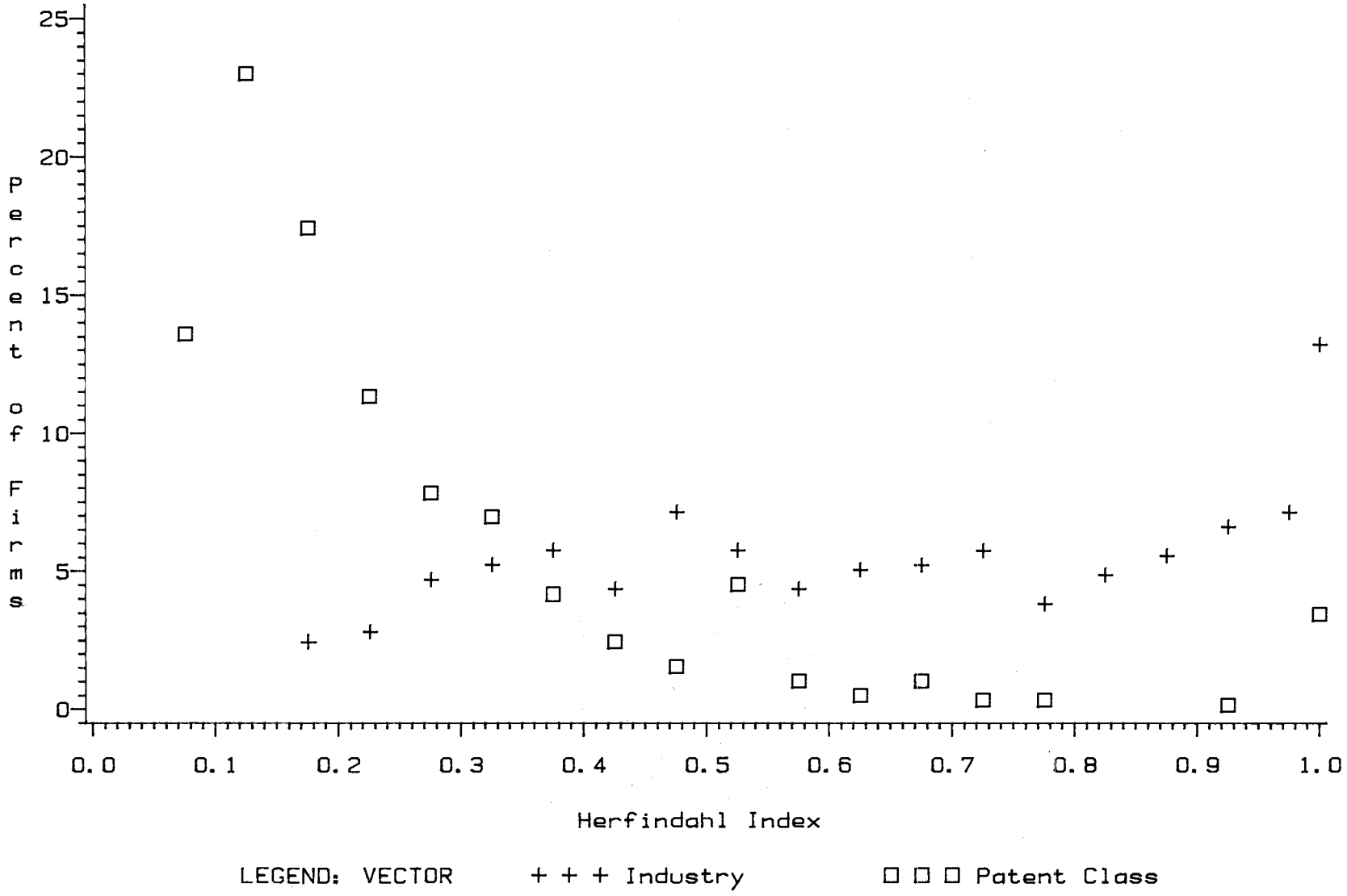
TABLE 2

STATISTICS FOR INDUSTRY GROUPS

INDUSTRY	SALES (billions)	NUMBER OF FIRMS: WITH ANY SALES	NUMBER OF FIRMS: WITH THIS AS PRIMARY INDUSTRY
FOOD & KINDRED PRODUCTS	70.2	92	46
TEXTILES & APPAREL	11.8	78	19
CHEM EXC. DRUGS	89.2	191	57
DRUGS & MED INSTRS.	23.5	66	28
PETROLEUM REF. & EX.	147.7	42	15
RUBBER & MISC PLASTICS	19.1	160	17
STONE, CLAY & GLASS	15.1	112	23
PRIMARY METALS	42.0	163	40
FABRIC. METAL PRODUCTS	30.5	245	37
ENGS, FARM & CONST. EQ.	28.5	112	23
OFFICE, COMP., & ACCTG	24.4	54	15
OTH MACH, NOT ELEC	25.0	263	49
ELEC EQ. & SUPPLIES	33.4	172	29
COMMUNICATION EQ.	30.0	142	34
MOTOR VEH & TRANS EQ	92.7	90	26
AIRCRAFT & AEROSPACE	32.7	84	19
PROF. & SCI EQ.	23.6	154	29
LUMBER, WOOD & PAPER	35.1	181	36
MISC MANU NEC	15.8	222	31

Figure One

Distribution of Hefindahl Indices of Patent Class and Industry Fraction Vectors



(139 industries) decreases it only to .31. I believe the difference is partly real, due to the tendency for even narrow product lines to incorporate several technologies. It is also partly an artifact of the data. Patents are a much less lumpy observation than establishments, so it is much easier to have a smattering of patents in many categories than it is to have a smattering of establishments in many industries.

Turning to the pool variable, this is constructed using the technological position vectors and some measure of the R&D of each firm. As discussed below, for some purposes we want the pool of R&D annual flows, and for some we want the pool of accumulated R&D stocks. If, for example, r_j is the 1976 R&D expenditure of firm j, then the 1976 R&D potential spillover pool for firm i is:

$$s_i = \sum_{j \neq i} P_{ij} r_j \qquad (2)$$

where P_{ij} is the technological proximity of i and j as defined in (1).[7] Note that Spence's θ has been omitted from Eq. (2). As long as θ is constant across firms, this means only that the units of the pool variable are arbitrary. Since I will be estimating elasticities, this is of no concern. To the extent, however, that θ varies across firms, the pool variable is not measured correctly.

A within-cluster spillover pool is calculated in the same manner,

---

7. In empirical work below, the variables $f_{i1} \ldots f_{iK}$ and s_i will be used as regressors in the same equation. Therefore, it may be useful to note the relation between s and the f's in matrix terms. Define F as the 573x49 matrix whose rows are the f_i's. Let F_N be the matrix derived from F by normalizing each row so its sum of squares is unity. Then the column vector s is given by $s = (F_N F_N' - I)r$ where r is the vector of firms' R&D spending.

but with the summation running only over firms in the same cluster. The procedure for clustering the firms into technological groups is described in Appendix B. Basically, I assume that the distribution of each firm's patents over classes is generated by a stable multinomial distribution; I group firms so that the likelihood is high that the patents of firms in the same group could have come from the same underlying multinomial distribution. This approach yields an iterative clustering algorithm that simultaneously estimates the underlying multinomial distributions for each group and assigns firms to groups. The resulting groups are shown with some relevant statistics in Table Three. Table Four presents a cross-tabulation of firms by these groups and industries. For this purpose, the PICA industry distribution data was ignored, and the firms were assigned to industries on the basis of their Compustat primary SIC.

For these pools to be reasonable measures of the potential R&D spillovers, it must be the case that the firms in the sample account for most of the relevant R&D. The amount of R&D done by out of sample firms is, of course, difficult to quantify, but two relevant comparisons can be made. First, the sample firms did 94% of the R&D reported by all Compustat manufacturing companies in 1976. Second, the current sample reported R&D equal to 86% of that reported to NSF in its annual survey in manufacturing product areas. Table Five presents a comparison by the 19 industry groups of the R&D of the sample firms and the R&D reported to the NSF in the corresponding product field. For this purpose, each firm's R&D was allocated to

TABLE 3

VARIABLE MEANS FOR TECHNOLOGICAL CLUSTERS (1976)

CLUSTER	NUMBER OF FIRMS	SALES	EMPLOYEES	PHYSICAL PLANT	R&D	PATENTS	TOTAL SPILLOVER POOL	IN-CLUSTER SPILLOVER POOL
ADHESIVES & COATINGS	30	555	12.6	395	4.6	4.7	2052	52
CHEMISTRY,CARBON	45	1871	28.1	1338	40.6	63.8	4079	1265
CHEMISTRY,ELECTROCHEMISTRY	16	1149	26.4	695	56.3	70.4	3615	560
CHEMISTRY,ORGANIC	21	946	19.6	391	42.4	52.6	3044	673
CLEANING AND ABRADING	16	817	22.1	214	9.3	14.5	3779	68
COMPOSITIONS	20	8623	29.3	5887	38.8	77.9	3740	538
CUTTING	24	212	4.3	88	2.5	4.6	2319	23
ELEC COMPUTERS & DATA PROCESSING	21	1538	36.1	784	80.1	59.3	3856	1065
ELEC TRANSMISSION & SYSTEMS	34	924	23.8	316	22.7	49.8	3746	529
ELECTRONIC COMMUNICATION	28	1430	39.2	530	36.1	55.1	3989	669
FLUID HANDLING	27	1003	21.1	660	16.3	32.1	4114	229
FOOD	34	1441	22.8	435	8.1	5.3	1907	200
MEASURING, TESTING & SIGNALING	31	556	12.0	143	14.4	12.7	3672	202
MEDICAL	13	571	14.0	195	16.5	18.5	2379	103
METALS AND METAL WORKING	33	1441	26.2	1099	15.5	15.0	3711	275
MISC CONSUMER GOODS	24	507	12.9	157	6.7	8.8	1919	86
POWER PLANTS (NON-ELECTRIC)	51	2729	49.0	732	70.7	40.0	4419	2131
RECEPTACLES & PACKAGES	24	575	10.1	413	4.6	8.4	1971	66
REFRIDGERATION & HEAT EXCHANGE	29	578	10.7	207	6.3	11.8	3180	95
STATIC STRUCTURES	27	289	5.9	152	1.5	3.5	2221	17
VEHICLES	25	737	13.1	258	12.9	15.0	3144	172
OVERALL	573	1379	22.3	698	25.3	29.8	3264	528

Note: Sales, Physical plant, R&D and Pools are in millions of dollars. Employees is in thousands.

TABLE 4

DISTRIBUTION OF FIRMS BY TECHNOLOGICAL AND INDUSTRY CLUSTERS

INDUSTRY GROUP / TECHNOLOGICAL GROUP:	1 ADHESIVES & COATINGS	2 CHEMISTRY, CARBON	3 CHEMISTRY, ELECTROCHEMISTRY	4 CHEMISTRY, ORGANIC	5 CLEANING AND ABRADING	6 COMPOSITIONS	7 CUTTING	8 ELEC COMPUTERS & DATA PROC	9 ELEC TRANSMISSION & SYSTEMS	10 ELECTRONIC COMMUNICATION	11 FLUID HANDLING	12 FOOD	13 MEASURING, TESTING & SIGNALI	14 MEDICAL	15 METALS AND METAL WORKING	16 MISC CONSUMER GOODS	17 POWER PLANTS (NON-ELECTRIC)	18 RECEPTACLES & PACKAGES	19 REFRIGERATION & HEAT	20 STATIC STRUCTURES	21 VEHICLES	TOTAL
1 FOOD & KINDRED PROD.		3		1	1	2						29			1			1	2			40
2 TEXTILES & APPAREL	13	3							1					3						1		21
3 CHEMICALS EXC. DRUGS		27	1	1	3	6	1	1		1	5		1	1		1	2	3		1		55
4 DRUGS & MEDICAL INSTR.		5		18							1	2	4	2								32
5 PETROLEUM REF. & EXT.		1				13					1									1		16
6 RUBBER & MISC. PLASTICS	1	5	1				2				5									1		21
7 STONE, CLAY & GLASS	2	1		3					1		8				1	1		1	1	2		21
8 PRIMARY METALS	1					1			6						14				1	1		24
9 FABRIC. METAL PROD.	1						6				1				3	3	5	4	4	4		31
10 ENGS, FARM & CONST. EQ.	1				1				1						1	1	6		2	1	10	24
11 OFFICE, COMP. & ACCTNG.			4					11		2			1							1		19
12 OTHER MACH, NOT ELEC.	2			1	5		9				3		2		4		11	1	5	2	3	48
13 ELEC. EQ. & SUPPLIES					1			1	17	3			5		2	1	2		4	2		38
14 COMMUNICATION EQ.			3					4	5	15			3						1		1	29
15 MOTOR VEH. & TRANS. EQ.							2	1		1			4		1		14		3		8	34
16 AIRCRAFT & AEROSPACE	1							1		2	1		4		2	1	5					17
17 PROF. & SCI. EQ.			8		1				2	2			6	5					2			26
18 LUMBER, WOOD & PAPER	7						1	1			2				1			14		5		31
19 MISC MAN. NEC	1								1			2	3			12			1	4	1	25
20 CONGLOMERATES			1		1		1	1		2		1	1		2	1	7			1	1	20
TOTAL	30	45	16	21	16	20	24	21	34	28	27	34	31	13	33	24	51	24	29	27	25	573

TABLE FIVE

COMPARISON OF R&D TOTALS FOR SAMPLE FIRMS AND NSF SURVEY

INDUSTRY	NSF*			573 firm
	Total	Federal	Company	Sample
FOOD & KINDRED PRODUCTS	329	-	-	562
TEXTILES & APPAREL	82	-	-	92
CHEM EXC. DRUGS	1926	-	-	1714
DRUGS & MED INSTRS.	1091	-	-	969
PETROLEUM REF. & EX.	767	52	715	610
RUBBER & MISC PLASTICS	502	-	-	313
STONE, CLAY & GLASS	263	-	-	231
PRIMARY METALS	506	26	481	478
FABRIC. METAL PRODUCTS	358	36	322	476
ENGINES AND MACH NEC	1085	23	1062	1100
OFFICE, COMP., & ACCTG	2402	509	1893	1303
ELEC EQ. & SUPPLIES	3125	1462	1663	760
COMMUNICATION EQ.	2511	1093	1418	895
MOTOR VEH & TRANS EQ	2778	383	2395	2251
AIRCRAFT & AEROSPACE	6339	4930	1409	983
PROF. & SCI EQ.	1298	155	1144	1069
LUMBER, WOOD & PAPER	420	-	-	471
ALL OTHER	217	5	212	249
TOTAL	26,093	9186	16,906	14,513

*Source: NSF (1979)

industries in proportion to its (1972) sales. This, of course, under-allocates to R&D-intensive industries, and to those that grew faster than average 1972-76. The former is visible in the Table in the fact that Food and Textiles both show more R&D by these firms than was reported to NSF; the growth problem may partially explain the computer shortfall. Even allowing for these tendencies, there appears to be significant under-representation in the current sample in Computers, Electrical Equipment, Communications and Aircraft and Aerospace. This introduces a potential bias of unknown importance into the subsequent analysis.

III. Research Intensity

We test for technological position, industry and spillover effects in 3 areas: the distribution of R&D effort itself, the production of new knowledge, and the growth of total factor productivity. For R&D intensity, we estimate an equation which is an extension of that used in Bound, et al (1984):

$$\log(r_i)=\beta_1 \log(q_i) + \beta_2 \log(C_i) + \beta_3 M_i + \beta_4 \log(s_i^T) + \delta(s_i^C/s_i^T) + \sum_k \gamma_k f_{ik} + \sum_\ell \alpha_\ell g_{i\ell} + \varepsilon_{1i} \quad (3)$$

where r_i is the annual R&D of the i^{th} firm, q_i is its sales, C_i is its capital stock, M_i is its sales-weighted average market share, and s_i is the pool of spilled research potentially available to it.[8] The

8. Throughout this paper, the subscript k will refer to the technological areas and the subscript ℓ will refer to industries. Capital letters represent stocks, and lower case letters the related flows.

superscripts T and C refer, respectively, to the total and within-cluster pools. The f_{ik} and $g_{i\ell}$ are the patent class and industry fractions defined above. All the ε_i's in this paper are stochastic errors assumed independently but not necessarily identically distributed.

This equation is not an R&D demand equation derived from optimal firm behavior; it is simply a device to examine patterns of R&D intensity descriptively. It is motivated by heuristic arguments about things that affect the cost and benefits of R&D. Since R&D is like a fixed cost, greater sales imply a greater return to R&D, at least to the extent that the external market for innovations is imperfect and innovation itself cannot be expected to increase sales dramatically. I include physical capital to allow for input complementarity, and for comparability with Bound, et al.

The firm's average market share is included on the notion that, in an environment of imperfect appropriability, the knowledge that your research may help your competitors is a disincentive to your doing R&D. Firms with large market shares may be less worried about this, and hence choose a higher R&D intensity.

The spillover pool affects R&D by increasing its productivity; this should cause the optimizing firm to choose a higher R&D intensity where the pool is greater.[9] The inclusion of the log of the total pool and the fraction that is within-cluster is an approximation to a

---

9. A more satisfactory story about the market share and pool effects requires explicit modelling of the firms' competitive interaction in the presence of spillovers. This will be a focus of future work.

specification where the "effective pool" is given by:

$$s_i^E=(1 + \delta)s_i^C + s_i^O = \delta s_i^C + s_i^T = s_i^T[1 + \delta(s_i^C/s_i^T)] \qquad (4)$$

where the superscript O refers to the out-of-cluster pool; δ is then the premium associated with the within-cluster portion. If $\delta(s^C/s^T)$ is small, then $\log(s^E)\approx\log(s^T) + \delta(s^C/s^T)$. For this sample, s^C/s^T has a mean of .15; δ turns out to be less than 1.

If variations in technological opportunity are a major factor in determining the cost of R&D, then we would expect the technological position effects in Eq. (3) to be important; that is, we should reject the hypothesis that the γ_k's are all equal. Similarly, if variations in demand are important determinants of the returns, this should be reflected in the industry effects (the α_ℓ's).

To estimate equation (3), a cross-section of firms for 1976 was chosen, 1976 being the year in which R&D was reported for the greatest number of firms. The definition of the R&D, sales, physical capital and market share variables is described in Appendix A. The pool variables are constructed using the weighted sum of other firms' reported 1976 R&D. In order to use the same sample in this and the next section, firms with no patents 1975-77 were dropped (though their R&D is still in the pool), leaving 537 firms.

The first column of Table 6 tests for market share and spillover effects without allowing industry or technological position effects. Both effects are positive and significant quantitatively and statistically. The cluster premium δ is also positive and significant. These effects remain, though the pool effect is diminished, when we add

TABLE 6

RESULTS OF R&D INTENSITY EQUATION ESTIMATION
(1976 Cross-Section)

Dependent Variable: Log of R&D

	1	2	3	4	5
Log(Sales)	.980	.900	.943	.877	.931
	(.091)	(.102)	(.095)	(.095)	(.041)
Log(Net Plant)	-.156	.083	-.087	.052	
	(.076)	(.090)	(.082)	(.084)	
Market Share	2.63	2.13	2.80	2.51	2.57
	(.715)	(.547)	(.563)	(.546)	(.525)
Log (Total Pool)	.749	.292	.391	.267	.276
	(.086)	(.097)	(.071)	(.080)	(.077)
Cluster Pool/Total Pool	1.08	.560	.066	-.062	
	(.255)	(.221)	(.324)	(.312)	
F-statistic on Overall Pool Effect	53.3	8.5	4.5	2.4	
	(2,531)	(2,513)	(2,483)	(2,465)	
F-statistic on Industry Effects		16.8		5.4	
		(18,513)		(18,465)	
F-statistic on Technological Area Effects			5.8	2.0	
			(48,483)	(48,465)	
R^2	.719	.823	.822	.853	.853
$\hat{\sigma}$	.901	.728	.752	.697	.696

Notes: 537 observations. Numbers in parentheses under coefficients are heteroskedasticity consistent standard errors calculated according to White (1980); F-statistics are not corrected for heteroskedasticity.

F critical values:	.95	.99
(2,400)	3.0	4.7
(20,400)	1.6	1.9
(50,400)	1.4	1.6

industry effects in Column 2; in addition, the industry effects are highly significant. Column 3 gives the results controlling for technological position effects but not industry effects. Again the pool variable is significant, but the cluster premium now goes to zero. The technological area effects are also significant themselves.

Column 4 allows for all effects. The pool coefficient is significant, with no evidence of a premium for the within-cluster portion.[10] The elasticity of own R&D with respect to the weighted sum of all others is between .2 and .3. It should be emphasized that the absence of a premium for the within-cluster firms says only that after weighting by proximity, there is no further differential to the in-cluster firms (who are proximate by construction).

These results imply a significant effect of the spillover pool at a given location in technology space, over and above the "pure" positional effect that is interpreted in terms of technological opportunity. There is no way of knowing, however, that the pattern of technological opportunity can be captured completely by a linear relationship with the position variables as specified here. Since the pool variable depends (non-linearly) on position, it could be picking up residual higher-order technological opportunity effects. If we were to allow for an arbitrarily complicated pattern of technological opportunity, we could not distinguish a separate effect of spillovers, even in principle, since any reasonable definition of the spillover

---

10. It should seem peculiar that the t-statistic for the pool coefficient is so much more significant than the F-statistic for both pool-related coefficients. This is due to the fact that the White correction, in this case, reduces the estimated standard error; the F-statistics have not been corrected.

pool must depend on technological position. Restricting technological opportunity effects to being linear in technological position should, therefore, be viewed as an untestable identifying restriction.

The market and technological position effects are also both significant, as theory would predict. Note that the difference in significance between the two is mostly due to degrees of freedom; the fraction of the variance explained by the market effects is only slightly larger. (See Table Ten and related discussion below.)

IV. Knowledge Production

The next area in which the effects of spillovers should be detectable is in the innovative output of the firm. We postulate a production function for new knowledge which, after taking logs, has the form:

$$\log(k_i)=\beta_1 \log(r_i) + \beta_2 \log(s_i^T) + \delta(s_i^C/s_i^T) + \sum_k \gamma_k f_{ik} + \sum_\ell \alpha_\ell g_{i\ell} + \varepsilon_{2i} \qquad (5)$$

where k_i is the new knowledge produced by the firm and the other variables are as above.[11] In practice, we use patents as an indicator for k.[12] Note that this formulation implies that spillovers increase the productivity of own R&D; below I allow them also to influence

---

11. Actually, we expect knowledge production to depend on a distributed lag of R&D, but this lag structure is difficult to identify, and much of the weight appears to fall on the contemporaneous R&D. See Pakes and Griliches (1984) and Hall, Griliches and Hausman (1983).
12. There has been a long debate on the general question of the usefulness of patent data as an indicator of inventive output. For present purposes, suffice it to say that patents have repeatedly passed tests of their economic relevance. See Schmookler (1966), Pakes and Griliches (1984), Bound, et al (1984), Pakes (1984), and Hirschey (1982).

TABLE 7

RESULTS OF KNOWLEDGE PRODUCTION FUNCTION ESTIMATION
(1976 Cross-Section)

Dependent Variable: Log of Average Patents Applied for, 1975-1977

	1	2	3	4
Log(R&D)	.748	.759	.713	.722
	(.027)	(.028)	(.025)	(.027)
Log(Pool)	.629	.569	.971	1.060
	(.080)	(.093)	(.077)	(.083)
Cluster Pool/Total Pool	-.172	.026	.706	.834
	(.268)	(.277)	(.253)	(.263)
F-statistic on Overall Pool Effect	26.2 (2,533)	16.3 (2,515)	23.3 (2,485)	26.3 (2,467)
F-statistic on Industry Effects		2.4 (18,515)		1.6 (18,467)
F-statistic on Technological Area Effects			2.2 (28,485)	1.9 (48,485)
R^2	.731	.752	.779	.792
$\hat{\sigma}$	.887	.867	.843	.834

Notes: 537 observations. Numbers in parentheses under coefficients are heteroskedasticity consistent standard errors calculated according to White (1980); F-statistics are not corrected for heteroskedasticity.

F critical values:	.95	.99
(2,400)	3.0	4.7
(20,400)	1.6	1.9
(50,400)	1.4	1.6

conventional output productivity directly. As above, the parameter δ is the premium for the in-cluster pool.

The interpretation of the technological position and industry effects in (5) requires comment. Differences in technological opportunity should again cause technological area effects to be important. In addition, there are significant variations in the "propensity to patent" different kinds of knowledge. This is because the value of patent protection differs for different technologies, as does the difficulty of successful application. This has the effect of associating differing amounts of knowledge and differing amounts of R&D with each patent. These variations will appear as technological position effects. A priori, we do not expect industry effects to be important here, but allow for them to see if the data are consistent with this view.

The knowledge production function (5) was estimated on the same 1976 cross-section of 537 firms. For k, we use the average patent applications 1975-1977, attempting to average out some of the noise in the patent data. The correct way to model the patent process is to take into account the integer nature of patents. This is done using Poisson-type models as in Hall, Griliches and Hausman (1983). To simplify, I ignore this complication and estimate the log-log regression This is why the zero patent observations had to be dropped.[13]

The columns of Table Seven are analagous to those of Table Six. The spillover effect is significant throughout. Interestingly, its

---

13 . There are relatively few of these firms, since 10 or more patents 1969-79 were required for initial inclusion in the sample.

magnitude is greater when we control for technological position and such control also results in a within-cluster premium that is positive and significant. The coefficients imply that the positive externality from spillovers outweighs any negative competitive effects; a 10% increase in the R&D of all firms would increase total patents by about 17%; 7% from the effect of their own R&D on their own patents, and 10% from the collective effect of the increased pool (ignoring the feedback effect which the previous estimation implies would induce additional increases in R&D).

At first glance, this may seem like an implausibly large spillover effect. I believe, however, that it is consistent with a reasonable picture of R&D as a partial public good. In terms of marginal products evaluated at the mean of the data, the coefficients imply a return of 1.7 patents per million dollars of own R&D, .05 patents per million dollars of other firms' relevant R&D if they are outside the firm's cluster, and .09 patents per million dollars of other firms' relevant R&D if they are in the same cluster.

Allowing for industry effects has little effect on the pool coefficient, and the industry effects are not significant after controlling for the other two. Thus we get the expected result on the unimportance of industries in knowledge production. The technological area effects are marginally significant when controlling for spillovers and industry effects.

V. Total Factor Productivity

The final area investigated is the conventional overall productivity of the firm. We postulate a Cobb-Douglass production function for the firm's output:

$$q_{it} = L_{it}^{\beta 1} \; C_{it}^{\beta 2} \; K_{it}^{\beta 3} \; (S_{it}^{E})^{\beta 4} \exp(\sum_{k} \gamma_k f_{ik}) \exp(\sum_{\ell} \alpha_\ell g_{i\ell}) \qquad (6)$$

where L_{it} is labor input and K_{it} is the stock of knowledge of the firm. S^E is the effective pool of the stocks of other firms' research, analagous to the pool of their research flows above. I take logarithmic time derivatives and convert this into an equation for growth rates:

$$(\dot{q}/q)_{it} = \beta_1 (\dot{L}/L)_{it} + \beta_2 (\dot{C}/C)_{it} + \beta_3 (\dot{R}/R)_{it} + \beta_4 (\dot{S}^E/S^E)_{it} + \sum_{k} g_k f_{ik} + \sum_{\ell} \alpha_\ell g_{i\ell} \qquad (7)$$

Thus, in this formulation, the technological position effects are interpretable as differing rates of exogenous technological progress. In a perfect world, we would again expect no industry effects to be important in this equation. In such a world we would be measuring true shifts in the production function, which should not be explained by industry assignments once we control for technological position. In reality, however, there are three reasons to allow for industry effects. First, to get a measure of real output we need price deflators; these are imperfect and the errors in them have a distinct industry pattern (Griliches and Lichtenberg, 1984). Second, since

input measures are not corrected for capacity utilization, changes in this affect measured productivity. We certainly expect the growth rate of capacity utilization to show an industry pattern. Finally, including industry effects should mitigate the simultaneity bias that may be present due to the influence of the (expected) growth rate of demand on R&D.

I postulate once again that the effective spillover pool is a weighted sum of the in-cluster and out-of-cluster pools. It turns out, however, that the approximation used above does not work empirically in this equation. Instead, we write:

$$S_i^E = S_i^C + \lambda S_i^O = S_i^C[1 + \lambda(S_i^O/S_i^C)]; \tag{8a}$$

$$(\dot{S}^E/S^E) \approx (\dot{S}^C/S^C) + \lambda(\dot{S^O/S^C}) \tag{8b}$$

where λ is the relative value of the out-of-cluster portion with the in-cluster portion given weight of unity, and the $\dot{}$ indicates the time derivative of the entire expression in parentheses. The approximation is now valid only if $\lambda(S^O/S^C)$ is small. Though S^O is larger than S^C, λ turns out to be essentially zero, so the approximation is valid.

There are two conceptual issues that must be addressed regarding K in Eq. (7). First, in most work along these lines, the R&D stock is substituted for the knowledge stock, giving R&D a role directly analagous to physical investment in the production of output. Here, however, such an approach would be inconsistent with the previous assumption that the growth of the knowledge stock does not depend on

R&D investment alone. A more consistent approach is to use the patent stock as an indicator of accumulated knowledge. Both of these approaches are utilized in the empirical work described below.[14]

The second issue relates to the interpretation of β_3, and is most clearly stated in the context where the R&D stock is used for K. If we were to estimate Eq. (7) as written, we would be assuming that the output elasticity of the R&D stock is constant across firms. If firms choose inputs optimally, this is inconsistent with the large observed variations in R&D intensity.[15] An alternative is to assume that the marginal product, not the elasticity, is constant (Griliches and Lichtenberg (1984)). That is:

$$\beta_3 \equiv (\partial q/\partial R)(R/q) \equiv \rho_K(R/q) \tag{9}$$

where the stock of research R has been used instead of K and the firm subscripts have been suppressed. If we substitute (8b) and (9) into (7) and note that the time derivative of R is r (neglecting depreciation):[16]

$$(\dot{q}/q)_{it} = \beta_1(\dot{L}/L)_{it} + \beta_2(\dot{C}/C)_{it} + \rho_K(r/q)_{it} + \beta_4(\dot{S}^C/S^C)_{it}$$

$$+ \lambda(\dot{S}^O/S^C) + \sum_k g_k f_{ik} + \sum_\ell \alpha_\ell g_{i\ell} + \varepsilon_{3i} \tag{10}$$

---

14. One might also argue that the pool variable should be based on the output (patents) rather than the input. There are two reasons to avoid that approach. First, the variations in the propensity to patent make the interpretation of patent spillovers difficult (Jaffe (1983)). Further, patents do, after all, have something to do with appropriation. It seems problematic to use them as a measure of the potential leakage.
15. R&D intensity varies across firms far more than does capital or labor intensity (Schankerman (1978)).
16. In applying this formulation to industry data, Griliches and Lichtenberg (1984) found that a zero depreciation rate yielded a better fit than any other.

Estimating (10) instead of (7) means assuming that it is the marginal product of R&D, not its output elasticity, that is constant across firms, as suggested by economic theory.

In the formulation that uses the patents for the growth in knowledge rather than R&D, it is less clear that this approach is superior. The marginal product of patents is not likely to be equal across industries, since their cost varies greatly, as discussed above. The constant elasticity approach may therefore be more reasonable a priori. In empirical work it was found that only the rate of return form yields significant coefficients, so that form is reported below.

The same issue, and the same ambiguity, arise with the pool variable. Here, there is no economic force whatsoever equalizing rates of return because this is an input over which the firm has no control. Also, since we do not know Spence's θ, the units of the pool variable are not really right. We choose therefore to stick with the conventional elasticity approach for this variable.

The TFP equations were estimated using logarithmic differences, 1972-1977. The pools of other firms' R&D stocks were calculated for each of these years. The construction of the R&D stocks is described in Appendix A. For the R&D version, ratios of R&D to sales were calculated as the average of the 1972 and 1977 ratios.[17] For the patent version, average patents 1975-1977 was divided by the average

17. This eliminated 146 firms that did not report R&D in one of those years.

TABLE 8

RESULTS OF SALES EQUATION ESTIMATION-R&D FORM
(Differences, 1977 - 1972)

Dependent Variable: Log(Deflated 1977 Sales) - Log(Deflated 1972 Sales)

	1	2	3	4
ΔLog(Employment)	.721 (.047)	.692 (.038)	.657 (.033)	.661 (.033)
ΔLog(Net Plant)	.037 (.045)	.127 (.047)	.136 (.037)	.152 (.038)
R&D/Sales	1.98 (.41)	1.45 (.46)	1.06 (.30)	1.34 (.43)
ΔLog(Cluster Pool Stock)	.041 (.049)	.098 (.051)	.089 (.039)	.099 (.041μ
$\Delta\left(\frac{\text{Out of Cluster Pool Stock}}{\text{Cluster Pool Stock}}\right)$	.00034 (.00029)	.00035 (.00028)	.00024 (.00024)	.00035 (.00022)
F-statistic on Overall Pool Effect	.3 (2,421)	1.9 (2,403)	1.2 (2,373)	2.6 (2,355)
F-statistic on Industry Effects		6.3 (18,403)		1.4 (18,355)
F-statistic on Technological Area Effects			3.0 (48,373)	1.1 (48,355)
R^2	.618	.702	.723	.742
$\hat{\sigma}$	.191	.172	.172	.171

Notes: 434 observations. Numbers in parentheses under coefficients are heteroskedasticity consistent standard errors calculated according to White (1980); F-statistics are not corrected for heteroskedasticity.

F critical values:	.95	.99
(2,400)	3.0	4.7
(20,400)	1.6	1.9
(50,400)	1.4	1.6

of deflated 1972 sales and deflated 1977 sales.

Table 8 presents the result of the R&D version. Again the columns are arranged in the same pattern. With no industry or technological area effects (Column 1), the pool effect is insignificant. Adding industry effects improves things greatly. The physical capital coefficient becomes more reasonable, and the pool coefficient becomes significant; the industry effects themselves are also significant. Adding the technological position variables has a similar effect on the other coefficients, and the technological effects are significant when added alone. When all three effects are allowed, the pool variable is marginally significant, but neither industry nor technological effects are significant after allowing for the other and the pool effect. The parameter λ is essentially zero throughout, indicating that only the within-cluster pool has a discernible affect on TFP growth.

The estimated R&D coefficient of about 1.3 (over 5 years) implies that each dollar of R&D yields an annual sales increase of about 25 cents. The interpretation of this figure is clouded because R&D employees and machines are also included in L and C. The R&D coefficient may be roughly a measure of the excess return to R&D (See Schankerman (1981)). Since, however, we are not controlling for materials, this 25 cents is not pure profit. If materials are roughly half of sales, this implies an excess return to own R&D in the neighborhood of 10-15%. Another way to evaluate the own R&D effect is to look at the implied elasticity of sales with respect to R&D, which, by (9), is $\rho_K(R_i/q_i)$. Using the estimate of 1.34 from column 4, we get

1.34/5=.268 for ρ_K in annual terms. Evaluating R_i/q_i at the sample means, this yields an estimate of the elasticity of .030.

The pool coefficient is directly an elasticity. It says that when all your neighbors increase their R&D by 10%, you can produce 1% more output with the same amount of capital and labor.

Table 9 presents the results for the sales equation using patents rather than R&D as the measure of new knowledge. Because patents measure new knowledge production only with a large amount of error, OLS estimates in this case would be severely biased. Therefore, we utilize 2SLS with the R&D to sales ratio and pool (flow) to sales ratio as instruments. The results are broadly similar to the R&D form. Here, the pool effect is significant throughout; again neither technological position nor industries is significant after controlling for the other.[18]

The patent/sales coefficient shows a slightly different pattern than the R&D coefficient. The latter fell when we controlled for technological position effects, a pattern consistent with the technological opportunity story, if R&D and productivity growth are both high where opportunity is high. The patent coefficient, however, increases when we add technological area effects. It is not clear what interpretation to give to this.

---

18. The instrumental variable method renders the usual F-test invalid, since the restricted and unrestricted sum of squares are no longer independent. We must resort to an asymptotic chi-square test, called the test of over-identifying restrictions by Hausman (1984).

TABLE 9

RESULTS OF SALES EQUATION ESTIMATION--PATENT FORM
(2SLS estimates using r/q and Pool/q as instruments for Patents/sales)
(Differences, 1977 - 1972)

Dependent Variable: Log(Deflated 1977 Sales) - Log(Deflated 1972 Sales)

	1	2	3	4
ΔLog(Employment)	.749 (.052)	.703 (.042)	.662 (.034)	.670 (.034)
ΔLog(Net Plant)	.034 (.047)	.133 (.049)	.131 (.038)	.144 (.038)
Patents/Sales	.504 (.309)	.357 (.326)	.783 (.355)	1.01 (.410)
ΔLog(Cluster Pool Stock)	.065 (.052)	.093 (.051)	.088 (.068)	.101 (.045)
$\Delta\left(\frac{\text{Out of Cluster Pool Stock}}{\text{Cluster Pool Stock}}\right)$	.00050 (.00031)	.00040 (.00030)	.00027 (.00026)	.00041 (.00024)
χ^2_2-statistic for Overall Pool Effect	2.8	4.8	3.8	5.4
χ^2_{18}-statistic on Industry Effects		84.0		25.2
χ^2_{48}-statistic on Technological Area Effects			120.5	55.2
Second Stage Statistics: R^2	.595	.692	.709	.719
$\hat{\sigma}$	.197	.176	.179	.181

Notes: 434 observations. Numbers in parentheses under coefficients are heteroskedasticity consistent standard errors calculated according to White (1982); χ^2-statistics are not corrected for heteroskedasticity.

χ^2 critical values:	DF	.95	.99
	2	6.0	9.2
	18	28.9	34.8
	50	67.5	76.2

If we take the patent coefficient from Column 4, we get an estimate of the marginal product of a patent of about 1.0 million dollars over 5 years, or about $200,000 in terms of the annual sales increase. By the same argument as above, this implies a return to patenting over and above the normal return to the R&D that produced it on the order of $100,000 per year per patent granted.[19] The patent elasticity evaluated at the sample mean for K_i/q_i is about .039.

VI. CONCLUSION

This paper attempts to address an old question -- market pull versus technology push -- and a newer question -- the importance of spillovers -- by thinking about the location of firms in technology and market space. I find that, when allowing for all 3 effects, both technological position and industry are significant in explaining R&D intensity; the technological effects are significant in explaining patent productivity, and we cannot distinguish between the two in explaining TFP growth (at 1% significance). In interpreting these statistical statements, it is important to remember that different degrees of freedom are involved in each case. For perspective, Table Ten shows the simple analysis of variance for these and the spillover effects. Note, for example, that the technological effects actually explain considerably more of the TFP residual than do the market effects. Overall the pattern of effects is roughly what we would

---

19. The zero depreciation assumption implicit in this formulation may be more problematic for patents than for R&D.

TABLE 10

ANALYSIS OF VARIANCE FOR

SPILLOVER, TECHNOLOGICAL POSITION AND INDUSTRY EFFECTS

Equation	Percent of Variance Explained By:			
	Spillovers	Technological Position	Industry	Overall R^2
R&D Intensity[1]				
When added first	12.0	45.2	45.1	.561
When added last	.5	8.7	10.6	
Patent Productivity[2]				
When added first	12.9	19.7	11.6	.328
When added last	9.7	13.3	4.2	
Growth of TFP[3]				
When added first	.2	30.1	24.5	.350
When added last	.2	10.1	4.8	

Notes: 1. log(R&D) minus log(sales) and log(netplant) each multiplied by their respective coefficients from Table 6 Column 4

2. log(patents) minus log of R&D times its coefficient from Table 7 Column 4

3. growth rate of sales minus the growth rate of employment, the growth rate of net plant and the R&D/Sales ratio, each multiplied by their respective coefficients from Table 8 Column 4

expect on the basis of theory, with the qualifications discussed above regarding price indices and capacity utilization in the TFP equation. Both technology and industry are important in explaining R&D intensity and measured TFP growth; in patent productivity the technological effects predominate.

If we were to view the patent equation in isolation, there is no way to distinguish whether the technological position effects represent differences in opportunity or in the propensity to patent. If, however, it were only the latter, we would not expect the effect to carry over to the other equations.

One question raised by the degrees of freedom issue is the sensitivity of the results to the fineness of classification used in constructing the patent and sales distribution vectors. Exhaustive sensitivity analysis on this point would be quite expensive, but I did make a few experimental attempts to find a 20 category patent breakdown and a 50 category SIC grouping. I found that using the former reduced the fit quite a bit, while using the latter improved the fit very little over the results reported herein.

I find evidence of the effects of spillovers in several aspects of the innovation process, with the positive externalities apparently outweighing any negative competitive effects. When R&D in a firm's vicinity increases, the firm does more R&D itself, it produces more patents per R&D dollar, and its productivity grows faster, even controlling for the increased R&D and patents. These effects are present controlling for both industry and technological position effects as measured by sales and patent distributions.

It is interesting, and perhaps surprising, that these spillover effects "travel" quite far in technology space. In both the R&D and patent equations, a significant portion of the spillover effect is generated by firms outside the receiving firm's cluster. Only in the TFP growth equation is the spillover effect apparently localized. It is not clear whether this difference is substantive or an artifact of the pool variable construction. For the TFP equation, it is the ratio of the 1977 to 1972 pool of R&D stocks that is relevant. For this variable, the variance of the total pool is only 6.5% of the within cluster pool. By contrast, for the pool of annual flows the variance of the total is twice that of the within-cluster portion. Thus it may simply be that the overall pool changes too slowly to discern its impact in this type of regression.

The strongest spillover effect was found in the patent equation. As noted above, the estimates imply that the majority of the patents produced by a general increase in R&D come from the spillover effect. There is one possible problem with interpretation in this equation that should be noted. Patents are an indicator of innovation, but they are also a tool of appropriation. It could be that when others do more R&D, the firm worries more about protecting its own results and hence takes out more patents without necessarily producing more innovation. While there is probably something to this explanation, it seems unlikely that it is the dominant effect. If it were, I would expect the within-cluster premium to be much greater. As it is, the out-of-cluster portion is typically contributing 2/3 or more of the

total effective pool. These "long distance" spillovers are likely to be in the nature of basic research results, rather than very specific developments. Such research would not seem likely to induce an increasing propensity to patent.

There are three logical next steps that will be explored in subsequent research. The technological relationships embodied in the TFP and patent equations will be combined with assumptions about competitive behavior to yield an R&D demand equation incorporating spillovers. This should permit estimation of the three equations as a consistent system. In addition, this paper did not utilize fully the panel nature of the available data. Both the Compustat data and the patent class data are available over time. This may permit a fuller exploration of the spillover effects.

Finally, we have assumed herein that the conditions of appropriability do not depend on industry or technological area. This is surely wrong. To get a really useful picture of the spillover phenomenon, it will be necessary to allow for variations in the appropriability environment. This is not, however, simply a matter of allowing the coefficients to vary across the sample. Rather, variations in the appropriability environment have to be incorporated in the construction of the pool variable, since what is relevant is the environment of the spilling firm, not the receiving firm.

With multiple caveats on the interpretation of the present results, it is still interesting to dramatize their implications for the importance of the spillover phenomenon. As a thought experiment, imagine generating somehow an exogenous increase of 10% in the R&D of

all manufacturing firms. Ignoring the feedback to even greater R&D implied by the R&D intensity equation, the estimates herein predict a 17% increase in patents as a result. This, in turn, would lead to a 1.8% increase in output with given capital and labor, partially from the direct effect and partly from the spillovers. Without spillovers at either stage we would have predicted only a 7% increase in patents and only a .3% increase in output.

References

Addanki, Sumanth, D. Body and A. Jaffe, "Documentation for Data Set SPV," mimeo (1983)

Bernstein, Jeffrey I., and M. I. Nadiri, "Research & Development, Spillovers and Adjustments Costs: An Application of Dynamic Duality at the Firm Level," mimeo (1983)

Bound, John, C. Cummins, Z. Griliches, B. Hall, and A. Jaffe, "Who Does R&D and Who Patents?," in Griliches, ed, R&D, Patents and Productivity, University of Chicago (1984) (hereinafter Griliches (1984))

Cummins, Clint, B. Hall, E. Laderman, and J. Mundy, "The R&D Master File Documentation" mimeo (1984)

Griliches, Zvi, "Issues in Assessing the Contribution of R&D to Productivity Growth," BJE (1979)

_____, and F. Lichtenberg, "R&D and Productivity at the Industry Level: Is There Still a Relationship?" in Griliches (1984)

_____, "Interindustry Technology Flows and Productivity Growth: A Reexamination," REStat (1984a)

Hall, Bronwyn, Z. Griliches and J. Hausman, "Patents and R&D: Searching for a Lag Structure," NBER Working Paper #1227 (1983)

Hartigan, J.A., Clustering Algorithms, John Wiley and Sons, New York (1975)

Hausman, Jerry A., "Specification and Estimation of Simultaneous Equation Models, Griliches and Intrilligator, eds., Handbook of Econometrics (1984)

Hirschey, Mark, "Inventive Output, Profitability and Economic Performance," University of Wisconsin Graduate School of Business Working Paper # 5-82-28 (1982)

Jaffe, Adam, "Using Patent Data to Measure Technological Proximity and Research Spillovers Among Firms," NBER Summer Institue Paper (1983)

McQueen, J.B., "Some Methods for Classification and Analysis of Multivariate Observations," Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (1967)

Myers, S. and D. G. Marquis, Successful Industrial Innovation, National Science Foundation (1969)

National Science Foundation, Research and Devlopment in Industry, 1977, Surveys of Science Resources Series, Publication 79-313 (1979)

Pakes, Ariel, "Patents, R&D and the Stock Market Rate of Return," NBER Working Paper #786 (1982)

_____, and Z. Griliches, "Patents and R&D at the Firm Level: A First Look," in Griliches (1984)

Reinganum, Jennifer, "Dynamic Games of Innovation," JET (1981)

Rosenberg, Nathan, "Science, Invention and Economic Growth," Economic Journal (1974)

_____, Inside the Black Box: Technology and Economics, Cambridge University Press (1983)

Schankerman, Mark, "Essays in the Economics of Technical Change: The Determinants Rate of Return and Productivity Impact of Research and Development," PhD Dissertation, Harvard University (1978)

_____, "The Effects of Double-Counting and Expensing on the Measured Returns to R&D," REStat (1981)

Scherer, F.M., "Firm Size, Market Structure, Opportunity and the Output of Patented Inventions," AER (1965)

_____, "Demand-Pull and Technological Innovation: Schmookler Revisited," J Ind Ec (1982a)

_____, "Inter-Industry Technology Flows and Productivity Growth," REStat (1982b)

Schmookler, Jacob, Invention and Economic Growth, Harvard University Press, Cambridge (1966)

Shesko, Marilyn, "Program for Industry and Competitive Analysis," Harvard Business School mimeo (1981)

Spence, A. Michael, "Cost Reduction, Competition, and Industry Performance," EMA (1984)

White, Halbert, "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," EMA (1980)

_____, "Instrumental Variables Regression with Independent Observations," EMA (1982)

APPENDIX A: VARIABLE CONSTRUCTION

I. INDUSTRY DEFINITIONS

		SIC
1	FOOD & KINDRED PRODUCTS	20
2	TEXTILES & APPAREL	22,23
3	CHEM EXC. DRUGS	28 exc. 283,2844
4	DRUGS & MED INSTRS.	283,2844,384
5	PETROLEUM REF. & EX.	29
6	RUBBER & MISC PLASTICS	30
7	STONE, CLAY & GLASS	32
8	PRIMARY METALS	33
9	FABRIC. METAL PRODUCTS	34, exc 348
10	ENGS, FARM & CONST. EQ.	351,352,353
11	OFFICE, COMP., & ACCTG	357
12	OTH MACH, NOT ELEC	35,NEC
13	ELEC EQ. & SUPPLIES	365,366,367
14	COMMUNICATION EQ.	36,NEC
15	MOTOR VEH & TRANS EQ	37,NEC
16	AIRCRAFT & AEROSPACE	372,376
17	PROF. & SCI EQ.	38,NEC
18	LUMBER, WOOD & PAPER	24,26,27
19	MISC MANU NEC	348,21,25,31,39,NEC

II. OTHER VARIABLES

Sales: Deflated sales was calculated from Compustat sales figures using firm-specific price deflators, constructed as weighted averages of price series at the 4-digit SIC level. These price deflators are from the Bureau of Industrial Engineering of the U.S. Commerce Department. The firm weights are based on 1978 Compustat Business Segment Data. For more detail, see Addanki, et al (1983)

R&D: Taken from Compustat, with some addition of data from 10K reports where Compustat data was missing. See Cummins, et al (1982)

Net Plant: An estimate of the inflation-adjusted depreciated capital stock of the firm. It is based on the book value of net plant times the ratio of an investment deflator for the year to the investment deflator n years ago, where n is an estimate of the average age of the firm's capital. For details, see Cummins, et al (1982)

R&D Stocks: The R&D stocks used to construct the pool of stocks were calculated from R&D flows assuming 15%/year depreciation, and using as a starting value the first reported flow value divided by .22 (15% depreciation plus an assumed 7% annual growth in the flow since t=0). See Addanki, et al (1983)

Market Share: The PICA database gives market shares for each 4-digit SIC in which a firm has sales. To form the firm average, these were weighted by the fraction of the firm's sales going to the SIC.

APPENDIX B:

DESCRIPTION OF THE CLUSTERING PROCEDURE

Clustering is a generic term for a variety of techniques for grouping observations on multivariate data. The objective is to arrange the data in groups so that, in some sense, the observations are similar to other observations in the same group and dissimilar to all others. The easiest way to think of this is to visualize the observations in multi-dimensional space, and imagine grouping them so that the distances from the group means are minimized.

For many observations and even a few potential groups, the number of possible clusterings is a big number--not big like the federal budget deficit, but big like the number of atoms in the solar system. This fact, plus the fact that the objective function is grossly non-concave with respect to rearrangements, means that no clustering procedure exists that is guaranteed to find the best clustering for any real data. All procedures are somewhat ad hoc, results are sensitive to initial conditions, and no meaningful statistical tests of the resulting clustering can be performed. Clustering is a data description or simplification tool, not an analytic one. Its usefulness rests on facilitating subsequent analysis.

In the present application the goal is to identify groups of firms doing research in similar areas. The data to be used for this purpose is the distribution of the firms' patents over classes. The key attribute of these data is that they can be viewed as being generated by some

(unknown) underlying multinomial distribution: for each patent granted, the firm takes a draw from this distribution to determine what class the patent will be in. This particular structure of the data makes it possible to construct a clustering procedure with an explicit probabilistic basis. The clustering problem is the identification of groups of firms whose patents are generated by the same distribution; firms in different clusters draw their patents from different distributions. These underlying distributions must be estimated as we do the clustering, giving the procedure a bootstrap flavor.

The procedure utilized here is derived from something called the k-means algorithm in the clustering literature (Hartigan (1975), McQueen (1967)). In the k-means algorithm, observations are somehow grouped into k initial groups; the k means of these groups are calculated, and firms are then reassigned to the group such that the Euclidean distance to the group mean is minimized. Here, the multinomial nature of the data is exploited to derive an assignment criterion other than Euclidean distance. Instead of calculating group means, we calculate the fraction of all patents granted to firms in that cluster that occurred in each category:

$$\pi_{jk}=\left(\sum_{i\varepsilon j} n_{ik}\right)/\left(\sum_{i\varepsilon j}\sum_{k} n_{ik}\right) \qquad \text{(A1)}$$

where j indexes clusters, i indexes firms, and k indexes categories. π_{jk} is the probability of a patent issued to a firm in cluster j being in category k; n_{ik} is the number of patents issued to firm i in category k. On the provisional assumption that all firms in cluster j

derive their patents from the same multinomial distribution, these fractions are the maximum likelihood estimates of the parameters of the distribution.

We now consider reassigning the firms to other clusters. For each firm/cluster combination, we calculate the probability of the firm's observed distribution having come from the estimated multinomial parameter vector for the cluster. This probability is:

$$p_{ij}=\left[n_{i}.!/\prod_{k} n_{ik}!\right]\prod_{k}(\pi_{jk})^{n_{ik}} \qquad (A2)$$

where $n_{i}.$ is firm i's total patents. For the next iteration, the firm is assigned to the cluster where this probability is greatest. Fortunately, this can be simplified by noting that the term in brackets does not depend on j, and hence is irrelevant to the maximization and can be ignored. Also, maximizing $\log(p_{ij})$ is the same as maximizing p_{ij}, so what we really do is to choose j for each i to maximize:

$$\log(\tilde{p})_{ij}=\sum_{k} n_{ik}\log(\pi_{jk}) \qquad (A3)$$

When this is finished, we recalculate the vectors π_{j} using the new assignments; then we begin again. The process is continued until no reassignments occur.[20]

Using this procedure, it is natural to develop a measure of goodness of fit based on the log likelihood of the data conditional on the

---

20. Eq. (A3) is not defined for $\pi_{jk}=0$. In practice, if π_{jk} is zero and n_{ik} is not, we set p_{ij} to zero, preventing assignment of i to cluster j. If π_{jk} is zero and n_{ik} is also, we set $n_{ik}\log(\pi_{jk})$ to zero.

final clustering. The actual log likelihood is a mess, because it includes the bracketed term in Eq. (A2). We can, however, calculate a likelihood ratio in which these terms all cancel out. If we define $\pi_{\cdot k}$ as the fraction of all firms' patents taken in category k, we can calculate the log probability that each firm's patents come from that grand multinomial distribution. If we sum these over all firms, and subtract this from the sum of log probabilities conditional on the clustering, the bracketed terms drop out, leaving:

$$LR=\sum_{ik}\sum n_{ik}\log(\pi_{jk}) - \sum_{ik}\sum n_{ik}\log(\pi_{\cdot k}) \qquad (4)$$

This "statistic" measures the improvement in the likelihood gained by considering the firms' patents as coming from the cluster multinomial distributions rather than one grand distribution. The word statistic is in quotes because it has no known distribution, and probably no knowable distribution. It is only a qualitative and comparative measure of the fit.

This multinomial clustering procedure has several problems. First, it is not guaranteed to converge. Since the firms are considered for reassignment one at a time, it is quite possible for a firm to be assigned to a cluster that is in fact worse than its current one, once all reassignments are taken into account. The process can wander about, repeatedly moving firms back and forth. This behavior was observed and will be discussed below. It is not even assured that the likelihood function will improve at each iteration, although behavior where this was not the case was not observed. Even when the

procedure does converge, there is no guarantee that it is to a global maximum. In practice, one can only try various initial conditions and pick the best fit found.

Finally, this procedure does not determine the number of clusters to be identified. One can only try various numbers; of course, the fit should always improve when the number is increased. It may be possible to calculate something like an adjusted R^2 that would measure the quality of the fit, controlling for the number of clusters. This has not been done. In this application, it was found that the convergence properties of the procedure deteriorated quite suddenly if the number of clusters was made too small, so the smallest number that worked reasonably well was used.

The initial conditions for the clustering are specified by a set of unique hypothetical multinomial distributions, as many distributions as clusters desired; the iterative process is begun by assigning firms to these vectors using the likelihood criterion. Two general methods of obtaining these starting distributions were used. One method was to use the actual distributions over particular groupings of the data itself, as in the McQueen k-means algorithm. The other method was to specify arbitrarily distributions designed to be "far apart" in technology space and yet expected to attract significant numbers of firms. The results from several attempts are summarized in Appendix Table One. The approach which seemed to work the best was to choose a subset of the patent categories themselves, specifically those that had many firms having that category as the firm's primary

one. For this subset of categories, vectors were formed with .95 as the probability for the category and .001 as the probability for all others.[21] This set of pseudo-unit vectors was used as the starting seeds. Thus, on the first iteration, each firm is assigned to the cluster corresponding to its primary category, if its primary category is one of the targets; if not, it is assigned to its second, third, fourth, or whatever is necessary. Then, the actual distribution vectors of these clusters are calculated, and away we go.

The question of the number of clusters was settled when the program refused to converge if asked to find less than about 20 clusters. The final run, which was used in the regressions described in the body of the paper, to be used below, resulted in 21 clusters, which are summarized in Table 3. The name of the cluster is the name of the patent category that served as the initial cluster seed. This clustering was arrived at by starting with 23 targets, and then dropping two clusters that were small and close to others. Having done this, the program was also run starting at the beginning with the corresponding 21 pseudo-unit vectors, and a very similar clustering resulted.

Although it is obviously difficult to tell, the final clustering seems to be doing what we want it to do: distinguishing firms' technological focus. For example, Adhesives and Coatings picks up a mix-

---

21. Zeros are avoided because no firm can be assigned to a cluster that has a zero probability in any category for which the firm has even a single patent. (See note 20 above.)

ture of textile and paper companies. A few oil companies get moved in with the chemical companies. Xerox, an "office equipment" company by SIC, is at the center of the Chemistry, Electrochemistry cluster, along with Kodak and Polaroid. Interestingly, the big auto companies end up in Power plants (i.e. engines), leaving the equipment makers like John Deere and International Harvester in the Vehicles cluster.[22]

A known problem with clustering techniques in general is that they are capable of finding clusters where none exist. To get some indication of whether these data are "truly" clustered, a simulation experiment was conducted. Fake patent vectors for each firm were generated, using the firm's true total patents, but distributing them over classes randomly based on draws from the grand multinomial distribution of all the data. That is, observations were generated that truly belonged in one cluster. The clustering program was run on this data, using as starting seeds those used in the 21 cluster final run. After 10 iterations, the procedure was still reassigning over 100 firms at each iteration. The likelihood ratio was about 6000 (compared to 81000 in the chosen clustering) and was not improving,[23] providing some evidence that, when the procedure convergences in a well-behaved manner, some true clustering of the data does exist.

---

22. A complete listing of the firms by cluster is available from the author.
23. The fact that this clustering yielded a LR of 6000 is an indication of why the usual chi-square tests do not apply.

APPENDIX TABLE ONE

CLUSTERING RESULTS FROM SEVERAL RUNS

Run	No. of Clusters	No. of Iterations	Total Reassign-ments	Convergence Achieved?	Likelihood Ratio
1	23	8	201	Yes	80113
2	23	7	193	Yes	81628
3	20	10	184	Yes	70206
4	23	8	297	Yes	51832
5	21	10	231	Yes	80136
6	5	10	283	No	50839
7	20	15	270	No	77107

Starting Values:

1. Pseudo unit vectors based on classes with most patents.

2. Pseudo unit vectors based on most primary class firms.

3. Industry category fractions.

4. Random group category fractions.

5-7. Same as 2