NBER TECHNICAL WORKING PAPER SERIES

THE R&D MASTER FILE DOCUMENTATION

Bronwyn H. Hall

Clint Cummins

Elizabeth S. Laderman

Joy Mundy

Technical Working Paper No. 72

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 1988

There were many other contributors to the creation of this dataset: Sumanth Addanki, David Body, John Bound, Phil Farrell, Zvi Griliches, and Adam Jaffe. This research is part of NBER's research program in Productivity. Any opinions expressed are those of the authors not those of the National Bureau of Economic Research.

NBER Technical Working Paper #72
December 1988

THE R&D MASTER FILE DOCUMENTATION

ABSTRACT

This document describes the panel of publicly traded United States manufacturing firms which was created and updated at the National Bureau of Economic Research from 1978 through 1988 within the Productivity Program. The panel consists of about 2600 large manufacturing firms with three to twenty-seven years of data each; the period covered by the sampling frame was 1976 through 1985, with data back to 1959 where possible. There are approximately 70 variables for each firm-year of data, consisting of income statement and balance sheet variables and the corresponding common stock data. The technological data available for these firms consist of R&D expenditures and patents granted, both by date of application and by granting date. The patents data are available only through about 1981, due to the limitations of our sources and budget. The firms on the file are identified both by their CUSIP number and by name, making it feasible to match this data to other sources.

Bronwyn H. Hall
National Bureau of Economic Research
204 Junipero Serra Boulevard
Stanford, CA 94305

Clint Cummins
National Bureau of Economic Research
204 Junipero Serra Boulevard
Stanford, CA 94305

Elizabeth S. Laderman
National Bureau of Economic Research
204 Junipero Serra Boulevard
Stanford, CA 94305

Joy Mundy
National Bureau of Economic Research
204 Junipero Serra Boulevard
Stanford, CA 94305

CONTENTS

1. Introduction

The file described in this document consists of up to twenty-seven years of data for every U.S. manufacturing sector company which existed for three or more years sometime between 1976 and 1985 and was on one of Standard and Poor's Compustat files as of 1978-1985. This is a very comprehensive universe, consisting of the following firms:

1. Industrial - firms on the New York and American Stock Exchanges, large or actively traded Over-the-Counter firms.
2. OTC - the residual of the Over-the-Counter firms.
3. Full Coverage - approximately 2,000 additional OTC firms plus 1,300 non-NASDAQ firms or wholly owned subsidiaries.
4. Research - firms deleted from the Industrial File because they were aquired, merged, liquidated, or went private (no longer file with the SEC).

Companies are identified both by name and by CUSIP (Committee on Uniform Security Identification Procedures) number.

The structure of this document is as follows. First, we give a brief overview of the sample on the file. Then we list all the variables on the file in alphabetical order and describe how they were computed. In the final section, we give more detail on the construction of the dataset from the Compustat files and from data on individual U.S. patents from the Office of Technology Assessment and Forecasting.

Table 1.1 shows the provenance of the sample originally drawn in 1976; this is the sample for all versions of the file prior to March 1987. It was drawn from the following Compustat files: Industrial (1959-1978 and 1962-1981); Over-the-Counter (1961-1980 and 1961-1980); Research (1959-1978); and Full Coverage (1961-1980). The first column in the second row is the sample on the file and was obtained by

requiring that data exist in three continuous years (including 1976) on the following variables: book value of inventory, gross plant, net plant, book value of long-term debt, depreciation, and the market value of common stock. This last variable was ararely present on the full coverage file; many of the firms on that file are privately held or wholly-owned subsidiaries of other firms. For this reason, the Full Coverage file was used only for the sample through 1981. We found that many of these firms either were not separate entities or that they tended to move into our OTC and Industrial files during the period. We also found that they often filed SEC reports only occasionally, so there were many gaps in their data.

This original sample was augmented in the fall of 1986 and the spring of 1987 to include all manufacturing firms which existed on one of the files (except Full Coverage) for at least three years between 1976 and 1985. The additional firms and data came from the 1981-1985 Industrial files, the 1981-1985 OTC files, and the 1984 Research File (the 1985 file was unavailable). A fuller description of the construction of this dataset is in section 3.6. The bottom panel of Table 1.1 shows the final sample available after we combined all the data from the various sources.

TABLE 1.1

SAMPLE INFORMATION FOR THE COMPUSTAT DATA

January 1985 Panel

File	Years	Total no. of Firms	No. of Firms in Manufact. Sector	No. of Unduplicated Firms in 1976
Industrial	59-81	2500	1299	1294
Research	59-78	2500	414	138
Over-the-Counter	61-82	850	489	487
Full Coverage	61-80	3300	1008	781
All Files		4150	3210	2700

File	Years	No. in Our Sample*	Total no. of Observations	No. with Positive R&D
Industrial	59-81	1243	23,547	9583
Research	59-78	132	2092	665
OTC	61-82	443	5445	2909
Full Coverage	61-80	86	571	274
All Files		1904	31,655	13,431

October 1988 Panel

File	Years	Total no. of Firms	No. in Our Sample**	Total no. of Observations	No. with Positive R&D
Industrial	81-85	2400	393	5015	2859
Research	84	1500	5	73	46
OTC	81-85	850	231	1633	1279
Old Panel	59-81		1972	36,986	20,982
All Files			2601	43,707	25,166

* Three years of continuous data.

** In the first three rows, this number excludes the overlap with the old panel; these are new firms only. In the fourth row, it includes 68 firms which were in the base sample but did not make it onto the panel due to insufficient years of data. When the 1981-1985 data was added, these firms became eligible.

The total sales of the firms in our sample in 1976 was 1.05 trillion dollars and their R&D expenditures were 15.8 billion dollars. For comparison, the total company R&D reported by the National Science Foundation survey of industrial research and development for 1976 was 17.4 billion dollars.[1] Excluding the nonmanufacturing and service industries would reduce this total by about one half billion dollars. Thus our coverage of R&D is not quite as comprehensive as NSF's. There are two reasons for this: we do not include in our sample privately held companies or other small firms which the NSF surveys. Also, some firms which we do include may not report R&D expenditures publicly although they do report them on the NSF survey, which is confidential. On the other hand, there are two reasons why we might expect our total to be higher than the NSF total: although we attempt to remove government-sponsored R&D from our R&D variable by inspection of annual reports in many cases, we may not have taken all of it out. Similarly, the NSF survey specifically excludes R&D done by a firm or its subsidiaries in a foreign country, while the Compustat database may inadvertently include this R&D due to the consolidated nature of the 10-K and annual reports.

We can also compare the total sales for our R&D-doing firms with that computed by NSF from the R&D survey. They report that in 1977 (a year later than our numbers) the net sales of manufacturing firms which

---

1. National Science Foundation, Research and Development in Industry 1978, NSF80-307. The survey described in this document is actually administered by the Bureau of the Census for the NSF.

performed R&D was about one trillion dollars. The total sales for those of our firms which had positive R&D in 1976 was 956 billion dollars.

Further detail on the approximately 2600 firms in this panel is shown in Table 1.2. In the first year of the sample, 1959, there are 431 firms, rising to 2295 in 1976 and falling again to 2009 by 1981, and 1568 in 1985. Each firm in the panel has a continuous block of data (no gaps) so that, for example, we can form a balanced time series-cross section of approximately 1300 firms for ten years from 1968 to 1977.

TABLE 1.2

DATA AVAILABILITY BY YEAR

	Ending Year											
	76	77	78	79	80	81	82	83	84	85	Total	Total # of Firms
Beg. Year												
59	13	18	35	3	3	22	13	25	32	267	431	431
60	17	13	28	2	4	16	19	13	14	127	253	684
61	2	4	6		4	3	3	2	1	33	58	738
62	6	5	12	2	3	8	7	7	13	51	114	849
63	1	2	10	1	4	4	11	5	8	34	80	931
64	3		7	1	1	1	8	10	3	25	59	987
65	4	3	10	2	1	5		10	20	28	83	1070
66	2		11			6	4	2	9	31	65	1138
67	2	6	2	1	2	3	3	9	4	259	291	1432
68	7	3	12	7	6	13	11	13	11	114	197	1627
69	6	3	7	1	1	6	3	2	5	51	85	1712
70	3	1	3	3	2	5	3	3	7	21	51	1766
71	3	6	6	6	5	6	7	5	7	52	103	1869
72	3	4	8	4	2	3	7	7	8	53	99	1969
73	3	2	3	8	1	3	1	4	11	53	89	2057
74	6	4	5	21	8	6	3	11	12	99	175	2232
75				4	4	2				13	23	2254
76			6	8	11	2	1	1	2	11	42	2295
77					2		1	2	2	5	12	2225
78							2		2	16	20	2172
79					1	5	2	4	4	34	50	2054
80						2		3	5	51	61	2038
81								3	4	27	34	2009
82								2	6	60	68	1956
83									3	36	39	1886
84									2	11	13	1756
85										6	6	1568
Total	81	74	171	74	65	121	109	143	195	1568	2601	2601

Each cell in the table shows the number of firms whose data begins in the row year and ends in the column year.

The last column shows the total number of firms whose data is available in the row year.

2. Dataset Description

The final dataset is on one file: RNDPANEL.SAS.OCT88, or RNDPANEL.MASTER.OCT88. Table 1.1 shows the number of firms and observations from each Compustat file. Except for the Full Coverage file, over 60 per cent of these firms have positive R&D in 1976. The binary IBM format version of this file has the following characteristics:

```
DSNAME=RNDPANEL.MASTER.OCT88
VOL=SER=volume ID
DCB=(RECFM=FB,LRECL=869,BLKSIZE=17380)
```

The format is integer, floating point, and four character string variables. In Fortran, this would be

```
(I8, 10I4, 7A4, 7A4, 2A4, A4, A1, 4I2, 62F12.5)
```

and in SAS,

```
(8.0  10*4.0  $28.  $28.  $8.  $5.  4*2.0  62*12.5)
```

There also exists a SAS dataset on tape called RNDPANEL containing these variables, which is preferred if you plan to use SAS to read the data. The missing value code on the fixed format version of the file is -99999., while on the SAS dataset the missing value code is the usual SAS code . (period/dot).

In Table 2.1 we show an alphabetical list of the variables on the file with a short description of each one. The table also shows the Compustat item number if the variable came directly from that file, the length of the variable, and its byte location on the Fortran file. Character variables are indicated by 'CHAR.' Many of the variables come directly from the Compustat files and further information can be obtained by reference to that Codebook. Those which are computed for this dataset are described somewhat more fully in the section following

the tables. All dollar values are in millions of current U.S. dollars, except for deflated R&D expenditures, which is in millions of 1972 dollars.

TABLE 2.1

VARIABLES ON THE R&D MASTER FILE
(Alphabetical Listing)

Compustat item#	Variable	Type	Length	Loc	Brief Description
129	ACQUIS	NUM	12	126	Acquisitions (from stmnt of changes)
	ADJ	NUM	12	138	(Curr.Liab.-Debt)-(Curr.Assts-Invns.)
	ADJDEP	NUM	12	150	Depreciation adjusted for inflation
27	ADJFACT	NUM	12	162	Adjustment factor for common stock
	ADJFACTF	NUM	12	174	Adj. factor for common (fiscal year)
	ADJINV	NUM	12	186	Inventories adjusted for inflation
	ADJTOT	NUM	12	198	TOTAL Defl. using age structure
	ADPRICE1	NUM	12	210	IND price deflator, adj for FYR
	ADPRICE2	NUM	12	222	Firm price deflator, adj for FYR
45	ADV	NUM	12	234	Advertising expense
	BKCAP	NUM	12	246	Book value of capital stock
9	BKDEBT	NUM	12	258	Book value of long-term debt
3	BKINV	NUM	12	270	Book value of inventory
12	BKPLNT	NUM	12	282	Book net plant value
1	CASH	NUM	12	294	Cash and short-term investments
	CIC	NUM	4	9	CUSIP issue number and check digit
	CNAME	CHAR	28	49	Company name
20	COMINC	NUM	12	306	Income available for common
41	COSTGOOD	NUM	12	318	Cost of goods sold
12	CURRASST	NUM	12	330	Total current assets
	CUSIP	NUM	8	1	CUSIP number for firm
14	DEPREC	NUM	12	342	Depreciation and amortization
74	DFRTAX	NUM	12	354	Deferred taxes
26	DIV	NUM	12	366	Dividends per share
	DIVF	NUM	12	378	Dividends per share (fiscal year)
N22	DLCOMP	NUM	2	118	LCOMP excludes employee benefits
N25	DLSEAS	NUM	2	120	Employees include seasonal figures
N1	DMERGE	NUM	2	122	Major merger dummy (CS footnote)
	DUP	NUM	4	13	Compustat code for duplicate files
29	EMPLY	NUM	12	390	Number of employees (1000s)
60	EQUITY	NUM	12	402	Common equity
	FILE	NUM	4	17	Compustat file code
	FILER	CHAR	5	113	Source Compustat file
	FORPAT	NUM	12	414	Patent applications by foreign subs.
	FORPATG	NUM	12	426	Patents granted to foreign subs.
	FYR	NUM	4	21	Month of fiscal year end
	GRATE	NUM	12	438	Gross rate of return
	GROCAP	NUM	12	450	Gross cap. stock adj. for inflation
7	GROPLANT	NUM	12	462	Gross plant

Table 2.1 (continued)

Compustat item#	Variable	Type	Length	Loc	Brief Description
	INAME	CHAR	28	77	Name of Industry (from SIC)
18	INCOME	NUM	12	474	Income before xtry items
71	INCTAX	NUM	12	486	Investment tax credit
37	INVCAP	NUM	12	498	Total invested capital
30	INVEST	NUM	12	510	Capital expenditures
42	LCOMP	NUM	12	522	Labor compensation
	NETCAP	NUM	12	534	Net capital stock adj. for inflation
52	NOLCF	NUM	12	546	Net operating loss carry forward
25	NOSHARES	NUM	12	558	Common shares outstanding (1000s)
	NPLANT	NUM	12	570	Net plant value adj. for inflation
	NRATE	NUM	12	582	Net rate of return
	OLDPNL	NUM	4	25	Dummy for firm-yr from 1981 panel
13	OPINC	NUM	12	594	Operating income before depreciation
	PATENTS	NUM	12	606	Patent applications
	PATENTSG	NUM	12	618	Patents granted
43	PENSION	NUM	12	630	Pension and retirement expense
24	PRICE	NUM	12	642	Price of common (calendar year)
	PRICEF	NUM	12	654	Price of common (fiscal year)
	PRICE1	NUM	12	666	IND price deflator
	PRICE2	NUM	12	678	Firm price deflator
47	RENTAL	NUM	12	690	Rental expense
46	RND	NUM	12	702	R&D expenditures in current $
	RNDEFLT	NUM	12	714	R&D expenditures deflated
	RNDFLAG	NUM	2	124	Code for R&D corrections
	RORC	NUM	12	726	Calendar yr rate of return to common
	RORF	NUM	12	738	Fiscal year rate of return to common
	RSTOCK	NUM	12	750	Stock of R&D capital
12	SALES	NUM	12	762	Sales in current $
	SHARESF	NUM	12	774	Common shares (fiscal yr, 1000s)
	SIC	NUM	4	29	Compustat 4-Digit Industry code
	SMBL	CHAR	8	105	Stock ticker symbol
34	STDEBT	NUM	12	786	Debt portion of current liabilities
	STK	NUM	4	33	Compustat stock ownership code
	TL	NUM	12	798	Long-term debt adj. for age structure
	TOTAL	NUM	12	810	Invest in uncons subs & intang & other
	VAL	NUM	12	822	Market value (D+E) in current $
	VCOMS	NUM	12	834	Value of common stock in current $
	VCOMS2	NUM	12	846	Alternate val of common in current $
	VPS	NUM	12	858	Value of preferred stock in current $
	XREL	NUM	4	37	Compustat code for industry group
	YEAR	NUM	4	41	Year of observation (data as of end yr)
	ZLIST	NUM	4	45	Exchange listing and S & P Index code

2.1 Details of the Variable Construction

ACQUIS -- Compustat data item #129, acquisitions (from Statement of Changes in Financial Position), the funds for or costs incurred in the acquisition of companies during the year (includes the net assets acquired).

ADJ -- Value of net short-term assets, equal to current assets (Compustat data item #4) less the value of inventories (BKINV), less current liabilities (Compustat data item #5) plus the value of short-term debt (STDEBT). Short-term debt is added back in because it is treated elsewhere, and because it is assumed that interest is being paid on it, which is not the case for the other short-term liablities. The short-term liabilities of a firm are composed of accounts payable, income taxes payable, accrued expenses and costs on contracts, dividends declared, employee benefits, customer deposits, and the current portion of long-term debt.[2] This variable does not include unfunded pension liabilities.

The short-term assets of a firm are cash and short-term investments, accounts receivable, inventories, prepaid taxes, estimated future tax benefits, security deposits, supplies and tools not in inventory, and prepaid expenses.

ADJDEP -- This year's depreciation adjusted for the effects of inflation. This variable is DEPREC deflated by the ratio of the GNP deflator for fixed nonresidential investment AA (see NPLANT for a definition of AA, average age) years ago to the current GNP deflator.

ADJFACT -- Compustat data item #27, the cumulative adjustment factor for common stock splits, stock dividends, etc. This factor, applied to the per-share data, converts such data into terms of the share units prevailing in the last year of data.

ADJFACTF -- like ADJFACT, but for end of fiscal year stock data (PRICEF, SHARESF, DIVF), from the quarterly Compustat.

ADJINV -- the value of the firms' inventories adjusted for the effects of inflation. The inventories are valued at cost (book) each year unless the firm has specified LIFO as one of

---

2. This is included in short-term debt but also exists as a separate item on the Compustat tape. There is potentially some additional information here on the long-term debt issued 19 years ago.

its methods of inventory valuation. The use of LIFO implies that the reported inventory valuation will be too low and an attempt is made to adjust for this, using the change in inventories from year to year to obtain a measure of how old the inventories are and inflating them accordingly.

The process is started by setting ADJINV equal to the book value in the first year. Then the value of the inventories for each succeeding year is calculated as follows: if there is an increase in book inventory from last year to this, last year's adjusted inventories (ADJINV) are inflated by the ratio of the inventory price index this year to the price index last year and then the increase (assumed to be in current dollars) is added. If there was a decrease, last year's ADJINV is inflated in the same way and then written down by the ratio of this year's book inventories to last year's. The firm uses more than one method of inventory valuation, the book values are combined with the adjusted values using proportional weights derived from the reported ranking of the methods. Each firm uses up to three inventory methods and they are ranked in order of importance on the tape. We use the following rule of thumb to determine the weights:

Reported Number of Inventory Methods	Rank of LIFO	% LIFO as weight
1	1	100.0
2	1	66.7
2	2	33.3
3	1	50.0
3	2	33.3
3	3	16.7

ADJTOT -- Investments in unconsolidated subisdiaries and intangibles plus other investments (TOTAL) adjusted for inflation. In order to adjust the reported values of these items for inflation, we take the same approach as we did for inventories. In the first year, the value of ADJTOT is simply set equal to its book value. In each succeeding year, the previous year's ADJTOT is inflated by the ratio of the deflators for fixed residential investment in the two years and then the change in the book value of the investments is added. If there was a decrease in book, last year's inflated ADJTOT is multiplied by the ratio of this year's book value to last year's. As in the case of inventories, this approximation becomes more and more accurate as time goes by, but possibly at a slower rate, owing to the older age of the assets making up the TOTAL variable on average.

ADPRICE1 -- the industry-specific price deflator, adjusted for fiscal year, from the SPV file.

ADPRICE2 -- the firm-specific price deflator, adjusted for fiscal year, from the SPV file.

ADV -- Compustat data item #45, advertising expense.

BKCAP -- Net book value of capital stock. This variable is defined as

BKCAP = BKPLANT + BKINV + TOTAL

BKDEBT -- Compustat data item #9, book value of long-term debt.

BKINV -- Compustat data item #3, book value of inventories.

BKPLANT -- Compustat data item #8, net book value of the firm's physical plant.

CASH -- Compustat data item #1, cash and short-term investments.

CIC -- the Compustat CUSIP issue number (two digits which identify whether security is a stock, bond, etc.) and check digit.

CNAME -- 28-character alphabetic name of the company (all caps).

COMINC -- Compustat data item #20, income available for common.

COSTGOOD -- Compustat data item #41, cost of goods sold.

CURRASST -- Compustat data item #4, total current assets.

CUSIP -- CNUM, the Compustat identifying number for the firm (up to six digits).

DEPREC -- Compustat data item #14, depreciation and amortization.

DFRTAX -- Compustat data item #74, deferred taxes.

DIV -- Compustat data item #26, dividends per share. ADJFACT must be used to make this variable comparable from year to year.

DIVF -- Dividends per share on a fiscal year basis. ADJFACTF must be used to make this variable comparable from year to year.

DLCOMP -- A dummy which is one when Labor expense (LCOMP) excludes employee benefits (Compustat footnote number 22).

DLSEAS -- A dummy which is one when the number of employees includes substantial (>10%) numbers of seasonal and/or part-time employees (Compustat footnote number 25).

DMERGE -- a dummy which is equal to one if the data for this firm in this year was obtained by adding data from a firm on the research file (for data before 1980). For data added after 1981, this variable is equal to 1 if Compustat reported that the data was affected by a merger or acquisition (footnote 1 equal to AA) and a 2 if Compustat reported a major merger (footnote 1 equal to AB)

NB. This variable was wrong on the 11/16/81 edition of the file but was corrected as of the November 1982 edition.

DUP -- Duplicate company code, which identifies companies carried on more than one Compustat file. The codes are

00	No duplicate file
81	Primary Industrial file and Canadian file
83	Tertiary Industrial file and Canadian file
84	Supplementary Industrial file and Canadian file
86	Full Coverage file, OTC file, and Canadian file
91	Full Coverage file and Primary Industrial file
93	Full Coverage file and Tertiary Industrial file
94	Full Coverage file and Supplementary Industrial file
96	Full Coverage file and Over-the-Counter file
98	Full Coverage file and Canadian file

EMPLY -- Compustat data item #29, number of employees. This is the number of company workers as reported to shareholders. It may be an average throughout the year or an end-of-year number; the latter is reported if both are given. It includes part-time employees and the employees of consolidated subsidiaries.

EQUITY -- Compustat data item #60, common equity (as reported).

FILE -- the Compustat file code, with the following meaning:

1	Primary Industrial File
2	Bank File
3	Tertiary File
4	Supplementary Industrial File
5	Full Coverage File
6	Over-the-Counter File
11	Primary Industrial File and in the S&P Industrials Index
13	Full Coverage File

FILER -- a code for the source of the data for this firm from the Compustat Files. The possible values are

IND	the 59-78 or old 6281 Industrial File
IND81	the 62-81 Industrial File
IND82	the 63-82 Industrial File
IND83	the 64-83 Industrial File
IND84	the 65-84 Industrial File
IND85	the 66-85 Industrial File
OTC	the 61-80 or old 63-82 Over-the-Counter File
OTC81	the 62-81 Over-the-Counter File
OTC82	the 63-82 Over-the-Counter File
OTC83	the 64-83 Over-the-Counter File
OTC84	the 65-84 Over-the-Counter File
OTC85	the 66-85 Over-the-Counter File
FCV	the 61-80 Full Coverage File
RES	the 59-78 Research File

FORPAT -- the number of patents applied for by foreign subsidiaries of the firm during the calendar year (these patents are also included in the PATENTS variable).

FORPATG -- the number of patents granted to foreign subsidiaries of the firm during the calendar year (this number is also included in the PATENTSG variable).

FYR -- the month in which the fiscal year ends for this firm and year of data. Fiscal years ending between January 1 and May 31 are treated as though they ended in the prior calendar year.

GRATE -- the gross rate of return to capital, defined as the ratio of gross cash flows to GROCAP. Gross cash flows are the sum of the income before extraordinary items (INCOME), depreciation (DEPREC), and interest income (INTRST) less an inventory valuation adjustment and an imputed income from short-term assets (the prime rate times ADJ).

GROCAP -- the gross capital stock adjusted for inflation. This variable is computed as

GROCAP = GPLANT + ADJINV + ADJTOT.

GPLANT is gross plant revalued as for NPLANT but with the gross book value of the plant as input. GPLANT itself is not included on the tape.

GROPLANT -- Compustat data item #7, gross book value of the firm's physical plant.

INAME -- the 28 character name of the 4-digit industry to which Compustat assigned this firm.

INCOME -- Compustat data item #18, income before extraordinary items and discontinued operations.

INCTAX -- Compustat data item #71, income taxes payable.

INVCAP -- Compustat data item #37, total invested capital.

INVEST -- Compustat data item #30, capital expenditures (gross investment). The amount spent for the construction and/or acquisition of property, plant, and equipment, including that of purchased companies (acquisitions).

LCOMP -- Compustat data item #42, labor and related expense including salaries, wages, profit sharing, payroll taxes, and employee benefits. (See DLCOMP also.)

NETCAP -- the inflation adjusted net capital stock. This variable is

computed as

NETCAP = NPLANT + ADJINV + ADJTOT

NEWPATS -- the total number of patents applied for by the firm during the calendar year, from the 1982 OTAF patents granted tape.

NEWPATSG -- the total number of patents granted to the firm during the calendar year, from the 1982 OTAF patents granted tape.

NOLCF -- Compustat data item #52, tax loss carry forward.

NOSHARES -- Compustat data item #25, the number of shares outstanding. This variable was wrong on all editions of the tape through August 1982, but is fixed as of November 1982. It is not valid for firms which were merged, as indicated by the MERGER dummy.

NPLANT -- the net value of the plant adjusted for inflation. This quantity is obtained by multiplying the book plant value by the ratio of the GNP deflator for fixed nonresidential investment in the current year to GNP deflator AA years ago.

AA is the average age of the plant and equipment for this firm which is deduced in the following manner: an average age series is obtained as the ratio of accumulated depreciation (gross plant minus net plant) to depreciation this year. This assumes straight-ine depreciation.

Then a length of life of the current plant and equipment is also computed as the gross plant divided by this year's depreciation and a five-year moving average is taken of this series to smooth it. This year's average age is then adjusted by the ratio of this year's length of life to the moving average. This has the effect of smoothing the age series slightly. If the average age exceeds nineteen years, it is set to nineteen years, since the deflator is only available back to 1939.

NRATE -- the net rate of return to capital, defined as the ratio of net cash flows less the inflation adjusted value of depreciation (ADJDEP) divided by the net capital stock (NETCAP).

OLDPNL -- A dummy which is equal to one if this firm-year came from the 1985 edition of the panel (data through 1981).

OPINC -- Compustat data item #13, operating income before depreciation.

PATENTS -- the total number of patents applied for by the firm during the calendar year. (See section 3.1 for a fuller explanation of the derivation of this variable.)

PATENTSG -- the total number of patents granted to the firm during the

calendar year.

PENSION -- Compustat data item #43, pension and retirement expense.

PRICE -- Compustat data item #24, end of calendar year stock price (changed from Compustat's "dollars-and-eighths" format to decimal).

PRICEF -- end of fiscal year stock price (changed from Compustat's "dollars-and-eighths" format to decimal) from quarterly Compustat.

PRICE1 -- industry price deflator, unadjusted, from SPV.

PRICE2 -- firm price deflator, unadjusted from SPV.

RENTAL -- Compustat data item #47, rental expense. All costs charged to operations for the rental of space and/or equipment.

RND -- Compustat data item #46, expenditures on research and development. The private (company-funded) expenditures on the development of new products and services including software but excluding customer of government-sponsored expenditures, exploration, engineering expense, inventor royalties, and market research and testing. Because of the importance of this variable for most of the topics in this study, when it was randomly missing for one year in the middle of a continuous sequence of reported R&D, we interpolated a value using the model described in Appendix 2. This model assumes that the logarithm of R&D expenditures evolves as a random walk, which is justified by the patterns of R&D spending actually observed. The interpolation affected less than ten percent of the firms.

Before use, this variable is generally deflated by the deflator shown in Table 2.2, which is a weighted average of an index of hourly labor compensation and the implicit price deflator in the non-financial corporate sector. The weights are 0.49 and 0.51 respectively, and the methodology is patterned on Jaffe (1972). The underlying data is from U.S. Department of Labor, Productivity Costs in Nonfinancial Corporations, various issues.

RNDEFLT -- R&D Expenditures in millions of 1972 dollars. This variable is RND deflated by a deflator which is a weighted average of the index of hourly labor compensation and the implicit price deflator in the non-financial corporate sector. The deflator is shown in Table 2.2.

RNDFLAG -- A flag for the corrections to the R&D figures. If the value is zero, no corrections other than looking up a missing number have been made. The other values possible are the following:

1 Customer-sponsored R&D removed
2 Engineering removed

3 Other redefinition of R&D, including interpolation
4 Engineering included
5 From the 157-firm sample (i.e., a Nadiri or other number)
6 Horwitz corrections for expensed/deferred R&D applied

RORC -- This quantity is the one-period (year) rate of return to holding a share of the company's common stock over the calendar year. The definition of the one-period rate of return at time t is the following:

$$q_t = (p_t - p_{t-1} + d_t)/p_{t-1}$$

where p_t is the price of the common stock at the end of the year and d_t are the dividends received per share during the year. Note that for this definition it is important to be careful that p_{t-1} measures the same share of the company as p_t.

RORF -- fiscal year rate of return to a share of common stock, from quarterly Compustat. For the approximately sixty percent of the firms whose fiscal year ends in December, this variable will be the same as RORC. For the others, the rate of return for the four quarters that coincided most closely with the firm's fiscal year was obtained from the CRSP quarterly stock files (Center for Research in Security Prices 1986). Note that because this file contains quarterly data, the variable is computed for the four quarters which span the fiscal year most closely. This means that the timing can be off by as much as two months in one direction and one month in the other.

RSTOCK -- the stock of R&D capital, constructed from the history of R&D investment using a perpetual inventory model with declining balance depreciation. See Appendix 2 for details.

SALES -- Compustat data item #12, net sales. This is the amount of actual billings to customers for regular sales completed during the period, reduced by cash discounts, trade discounts, and returned sales for which credit is given to customers. Interest and equity income from unconsolidated subsidiaries, non-operating income, and income from discontinued operations are excluded.

SHARESF -- The number of shares outstanding at the close of the fiscal year. This variable is the one corresponding to PRICEF, ADJFACTF, DIVF, and RORF.

SIC -- DNUM, the Compustat 4-digit Industry code. This variable is between 2000 and 3999, or 9997.

SMBL -- 8-character stock ticker symbol for this firm. Owing to errors in the industrial file we received for 1959-1978, this variable is garbage for some firms before 1979.

STDEBT -- Compustat data item #34, debt in current liabilities.

STK -- Stock ownership code from Compustat. The meaning is

0 Publicly traded company
1 Subsidiary of a publicly traded company
2 Subsidiary of a company that is not publicly traded
3 Company that is publicly traded but is not on a major exchange

TL -- The value of long-term debt adjusted for its age structure. Long-term debt is assumed to have been financed by 20-year bonds. Given a matrix of bond prices in year t for a bond due in year s and the distribution of the firm's long-term debt by the year incurred, it is possible to adjust the face value of the debt by the bond rates in any given year. We obtain such a matrix of prices from the Moody's corporate BAA bond price series given previously. The difficulty is to determine the age structure of each firm's debt. In the absence of any history of the firm prior to 1958, the approach taken is to assume that all firms have an age structure of debt which is the same as the aggregate age structure from the 1958 Survey of Current Business. This distribution is shown in Table 2.3.

After 1958, we attempt to build up each firm's own long-term debt distribution in the following manner: starting with the 1958 distribution, in each successive year the amount of new long-term debt issued equals the change in long-term debt from the previous year to this plus the amount retired this year, which is assumed to be equal to that issued 20 years ago. There are two sources of error in this computation:

a) The firm's age structure of debt may not be the same as the aggregate in 1958.

b) The bonds may not always be for a twenty year term.

Because of these sources of error, in any year the gross new debt issued may be negative. If this is the case, we set this contribution to zero and rescale the entire twenty year of long-term debt accordingly. Given the age distribution of book value of long-term debt for each of the twenty-three years in the and the matrix of bond prices we can now compute the market value of debt as

$$(1)\quad TL_t = \sum_{S=t-20}^{t-1} BV_s \cdot P_{st}$$

where BV is the book value distribution and P_{st} is the price in year t of a bond due twenty years from year s.

TOTAL -- book value of investments in unconsolidated subsidiaries and others and intangibles. Sum of Compustat data items 31,32, and 33.

VAL -- Market value of the firm, equal to the sum of the value of the preferred stock (VPS) the value of the common stock (VCOMS), the long-term debt adjusted for its age structure (TL), and the short-term debt (STDEBT), less the net short-term assets (ADJ).

VCOMS -- Value of the common stock at the close of the calendar year. The value is the price of the common stock times the number of shares outstanding at the close of the year. This computation is invariant to ADJFACT so long as the two variables come from the same year of data.

VCOMS2 -- Average value of common stock throughout the calendar year. The value is the average of the high and low price for the year times the number of shares outstanding.

VPS -- The value of the preferred stock, computed as the preferred dividends paid during the year divided by the preferred dividend rate for medium risk companies (from Moody). This yield series is shown in Table 2.2 for the years 1958 through 1985.

XREL -- a four-digit code that identifies the specific S&P industry group in which each company in an S&P industry is contained. See Compustat for details.

YEAR -- two-digit year for this observation, 59 through 85.

ZLIST -- the Compustat Exchange listing and S&P Index code.

TABLE 2.2: DEFLATORS AND BOND YIELDS

Year	Preferred Dividend Rate Medium Risk Companies	Bond Yield	Fixed Investment Deflator	R&D Deflator
1940		0.0475	29.1	
1941		0.0433	30.9	
1942		0.0428	33.8	
1943		0.0391	35.7	
1944		0.0361	36.6	
1945		0.0329	36.6	
1946		0.0305	39.9	
1947		0.0324	46.8	
1948		0.0347	51.3	
1949		0.0342	52.8	
1950		0.0324	54.3	
1951		0.0341	58.9	
1952		0.0352	59.9	
1953		0.0374	61.0	
1954		0.0351	61.4	
1955		0.0353	62.6	
1956		0.0388	67.0	
1957		0.0471	70.7	0.598
1958	.0514	0.0473	70.6	0.616
1959	.0499	0.0505	72.0	0.631
1960	.0518	0.0519	72.2	0.647
1961	.0482	0.0508	71.8	0.658
1962	.0481	0.0502	72.3	0.670
1963	.0469	0.0486	72.9	0.680
1964	.0467	0.0483	73.6	0.698
1965	.0460	0.0487	74.5	0.711
1966	.0503	0.0567	76.8	0.737
1967	.0534	0.0623	79.3	0.768
1968	.0583	0.0694	82.6	0.809
1969	.0662	0.0781	86.6	0.855
1970	.0770	0.0911	91.3	0.906
1971	.0711	0.0856	96.4	0.956
1972	.0703	0.0816	100.0	1.000
1973	.0729	0.0824	103.8	1.064
1974	.0837	0.0950	115.3	1.170
1975	.0847	0.1061	132.3	1.285
1976	.0792	0.0975	138.7	1.361
1977	.0779	0.0897	146.0	1.459
1978	.0824	0.0949	157.8	1.573
1979	.0911	0.1069	171.3	1.718
1980	.1060	0.1367	186.8	1.870
1981	.1236	0.1604	201.3	2.010
1982	.1253	0.1611	210.1	2.100
1983	.1102	0.1355	207.8	2.175
1984	.1162	0.1419	208.8	2.250
1985	.1844	0.1272	211.8	2.268
1986	.0900	0.1100	212.2	2.289

TABLE 2.3

AGE STRUCTURE OF CORPORATE DEBT IN 1958

Age in Years	Fraction of Debt
19	.019
18	.023
17	.023
16	.009
15	.010
14	.025
13	.046
12	.046
11	.048
10	.057
9	.047
8	.047
7	.054
6	.073
5	.068
4	.071
3	.071
2	.076
1	.095
0	.092

3. Construction of the Dataset

The construction of the original dataset of June 1982 proceeded in three stages. First, a cleaned panel of firms was created from the four Compustat files with market value and asset data adjusted for the effects of inflation. Then we used this list of firms and their subsidiaries to produce a complete list of firm names which we searched for in the patent files. Finally we aggregated the patents for each firm and produced a merged file with the Compustat data, the patents in each year, and some additional variables from the Compustat file which we had missed in the first round. An outline of the flow of data in constructing the dataset is shown in Appendix 3 (available on request). The circled items are data files and the names of the programs which created them are given on the lines joining each pair of files. We summarize the function of each program below:

1. RSRCHBIN and OTCBIN recoded the EBCDIC copy of the Research and OTC files to Compustat 360/370 binary format, for compatibility with the Industrial file.

2. B4EDIT copied the Compustat files, recoding the common stock prices and number of shares to be the market value, and deleting some extra trailer records on the file.

3. B4SELCT selected companies from the Research file which were to be merged with companies from the Industrial or OTC files.

4. B4FIX spliced 3 companies from the research file to their successor firms on the industrial file. These firms had undergone a name change and their CUSIPs had changed.

5. B4MERGE merged selected companies from the Research file with their corresponding acquiring company on the Industrial or OTC file.

6. SLEPIAN selected the manufacturing sector firms, chose companies with good data in 1976 and at least two neighboring years, corrected any gross plant numbers for which we had updates, computed various stock variables from the Compustat data, and wrote out only the variables wanted for further analysis. The format of the SLEPIAN output files has one year per record, rather than one firm per record

and not all years are there for all firms.

7. SLEPUPD updated the SLEPIAN files with any Sales, Employment, or R&D corrections which were found during the scan of data for "unreasonable" jumps. An R&D flag was created indicating the natu of the corrections and any duplicating firms were deleted.

Because the goal of this project is to produce a dataset in which changes within a firm over time may be investigated, a major concern in construction of this file was the quality of the time series for each firm. We attempted to find and correct problems of noncomparability of data for a firm over time in two ways. The first was to scan the major variables for each firm (R&D, gross plant, sales, and employment) with a computer program which looked for large jumps in the data from year to year and for missing values for these variables. The criteria we chose may be summarized as follows:

1. Gross plant, sales, employment, or common stock missing or zero.
2. R&D expenditures missing.
3. Change in gross plant from year to year greater than 30% and 2 million dollars in absolute magnitude.
4. Change in employment from year to year greater than 30% and 300 employees in absolute magnitude.
5. Change in R&D greater than 100% and one million dollars in absolute magnitude.

The annual reports and 10-Ks for the firms thus identified were looked up and corrected values for the variables recorded if they could be found. Many firms had large jumps because of mergers or acquisitions: some of these could be "corrected" or at least minimized by the merging described in the next paragraph, but others remain uncorrected and unflagged.

The other way in which we attempted to improve the time series for these firms was to merge firms backwards (retroactively) where we could. This could be done if the data for a firm which had been deleted because it

was absorbed in a merger appeared on the research file. We attempted to find all such firms by scanning the research file for every firm which appeared on it because of merger or acquisition. This list of firms was then looked up in Mergers and Acquisitions to see whether the acquiring firm was on one of our files. If it was, and the merger occured before 1976, we added the data for the firm on the research file to the data on the industrial (or OTC) file for the years prior to the merger or acquisition. If the merger occured after 1976, we declared the data on the industrial file for the years following the merger to be noncomparable (gross plant changed to missing) and hence deleted those years from the series. This procedure was carried out only in the case of mergers of reasonable size, i. e., greater than about 3% of gross plant of the acquiring firm.

Merging and correcting the data was done by the programs B4SELCT, B4FIX, B4MERGE, and SLEPUPD. SLEPUPD also corrected the R&D numbers in several ways; any corrected numbers are marked with the variable RNDFLAG which is described later in this documentation. Briefly, we attempted to remove customer-sponsored R&D and engineering expense where we could find redefined numbers in more recent annual reports, we supplied a few numbers from the earlier work on the 157 firm sample which had been obtained directly from the company involved, we replaced a few numbers because they were for amortized R&D expense (mostly smaller firms), and, in a few instances, we interpolated one year of R&D where there was a smooth series with one observation missing in the middle.

3.1 Matching the 1982 OTAF Patent Dataset to Compustat[3]

This section describes the methods used to merge annual patent data (by organization) with the company-based accounting data provided by Compustat, which we described in the previous section. This task was difficult for two reasons. First, the number of companies in the 2 datasets is large. There are over 66,000 patenting organizations and more than 2700 companies with accounting data. The number of steps involved in merging and the possibilities for errors to creep in also added to this "large numbers" problem. This difficulty suggested that much of the merging operation should be computerized. Second, the companies in the 2 datasets had to be matched "by name", as the identification numbers in the 2 datasets are quite independent. Since there are many ways to write down the name of a company (or the names of its subsidiaries which may also have patents), the task could not be completely computerized.

For the "parent" companies in the Compustat dataset, the following steps were performed:

1. A list of "child" companies, owned by the "parent" company was made and typed into a computer data file. The source for determining ownership was Directory of Corporate Affiliation (1976). The list included the parent company itself, divisions of the parent, and subsidiaries. Affiliates, foreign subsidiaries and foreign affiliates were included in the list when their financial data was consolidated with the parent company. In this way, the parent data matched will be highly compatible with the R&D data from the Compustat dataset. For the 2695 Compustat "parents", about 16,000 names were gathered, or about 6 per company. Subsidiary or "child" names which are not similar to the parent name account for 83% of all

---

3. This section was prepared by Clint Cummins.

names, and for 14% of matches to the patent organization names, so they are worth the effort.

2. Each name in this list was looked up in the list of 66,000 patenting organizations. This was done by a computer program which then produced a record of matched companies. Along with company names from both lists, the computer also recorded the company identification numbers from both sources. These 2 numbers could then be easily used (by the computer) to merge the actual patent and accounting data from the 2 sources. The matching method used by the computer program distinguishes between generic words such as INCORPORATED and more specific words which serve to better identify a company.

3. The computer's record of matched companies was checked by hand. Matches which were incorrect were removed from the record. The computer record included many companies which were only similar in name and were not exact matches. About 10% of the computer's matches were unquestionably correct, while 2% were questionable enough to require further investigation, and 88% were incorrect matches of similar names. Although removing the 88% incorrect matches by hand involved some work, it was a necessary precaution. If the computer had ignored more possible matches, some true matches would not have come to light. This is a consequence of both the complexity of matching corporate names and the incompleteness of the subsidiary list described in step 1. Fortunately, removing the incorrect matches was simplified by using the computer's text editor to check (and remove) the matches visually. Using the text editor also ruled out transcription errors, since no identification numbers had to be copied by hand or retyped. The end result of this step was a corrected list of matches.

4. The corrected list was checked against the master list of all matches for other companies. Any patent organizations that were "matched" to more than one Compustat company were investigated and the question resolved. Often this meant checking ownership of a company by looking at the 10K annual report of its parent company. At other times, the address of the patent organization was obtained by looking up a specific patent granted, and this would resolve the question. Once all Compustat companies had been matched and checked, the result was a nearly final master match list.

5. The final step in the merging task was to look for matches in the reverse direction. In steps 1-4, a Compustat company was looked up in the list of patenting companies. This produced 4460 matches, an average of 1.6 per Compustat company. In step 5, a list of 8096 patent companies which were not matched in steps 1-4 (and who have at least 5 patents in the 11-year period under consideration) was printed. It is worth noting that most of the patent organizations on this list were foreign companies, so the method of steps 1-4 was quite sound. Note also that of the 66272 patent organizations, 53716 are on neither list, so there are many small patent organizations which would complicate the matching task considerably if it were to be done completely by hand. This patent

company list was then checked by hand against the list of Compustat parent companies. This reverse match picked up patent companies which were ignored by the computer matching program, and also picked up companies which had been removed by mistake from the computer match record in step 3.

The 8076 unmatched patent organizations were looked up (by hand) in the 1981 Directory of Corporate Affiliation, and in the Compustat notebook's list of parent comany names. This process gave us some idea of how the unmatched patent organizations were distributed among several classes of firms.

1. For those with 26-35 total patents in the 11-year period:

   394 names

   119 foreign

   275 not foreign

      111 unidentified firms

      164 identified firms

         11 address questions (resolved by looking up actual patent)
         10 Compustat, but not in our SIC sample
         10 non-Compustat (usually nontraded firms)
          2 already being investigated
          3 acquired firms
          3 foreign-owned firms

       25 new matches

         11 matches missed previously for unknown reasons
         10 possible acquired firms
          3 missed due to 4 known weaknesses of computer matching program
          1 new listing in 1981 Directory not found in 1976 Directory

2. For those with 6 total patents in the 11-year period:

    1003 names

     361 foreign

     642 not foreign

       540 unidentified firms
       102 identified firms
        19 address questions (resolved by looking up actual patent)
        26 Compustat, but not in our SIC sample.

12 non-Compustat (usually nontraded firms)
1 already being investigated
1 acquired firms
6 foreign-owned firms

37 new matches

10 matches missed previously for unknown reasons
18 possible acquired firms
5 missed due to 4 known weaknesses of computer matching program
3 new listings in 1981 Directory not found in 1976 Directory
1 typographical error in key word of OTAF name

The four known weaknesses of the matching program are as follows:

1. Name beginning with initials, such as B.F. GOODRICH.

2. Use of UNITED STATES versus U.S.

3. Buried keyword in foreign name, such as DEUTSCHE TEXACO.

4. Buried keyword in division name, such as BLARFO, DIVISION OF EXXON CORP.

Due to the large number of unidentified firms, a rule was devised to select the more important unidentified firms for further investigation. If the unidentified (domestic) patent organization had more than 50 total patents in the 11-year period, or if it had at least 5 total patents in the years 1975 to 1977, it was investigated a second time. This cutoff rule resulted in a list of 907 firms. Some 330 of the 907 had already been resolved by hand, so this left 577 to be investigated in the second round. This additional investigation involved looking up the firm in the Dictionary of Obsolete Securities; it is referred to as the 577/907 list below.

The result of this process was a list of 5426 OTAF organizations matched to Compustat parents or children (subsidiaries and divisions). A file containing records with CUSIP number - OTAF number pairs was prepared, and included the vector of patents by application date. This file is named BYOTAF.D82 (27-MAY-1982), and is sorted by OTAF number.

These 5426 matches were also printed out by Compustat parent, in

descending order of total patents, as a "dictionary of the match". This file is named DICTIO.D82.

In addition to these matches, matches to foreign-owned firms and other non-Compustat firms have been recorded. Other unidentified firms that were investigated were also recorded. These recorded firms and the actual matches have been kept in a set of large files on the Harvard Science Center Vax named MASTER3.PS, MASTER4.PS, MASTER5.PS,MASTER6.PS, and MASTER7.PS. The contents of the files varies; some are sorted by CUSIP number or by OTAF number, and others have dummy records for Compustat companies with no matched patenting organization. Complete documentation for these files is in the files AAADIR.DOC, PANELFILE.DOC, and UPDATING.DOC. OTAF organization in the MASTER3.PS file has a 2-digit "SRC" code which indicates the job which matched it, or what kind of non-match identification was made. These codes were used to select the BYOTAF.D82 matches from the MASTER.PS files. The codes are as follows:

Included in the BYOTAF.D82 merge file

1-22 computer-assisted matches from various prepared files of company and subsidiary names.

29,40 computer-assisted matches to Compustat firms not in our current sample.

30 "by hand" matches of unidentified firms with 35+ patents

31 "by hand" matches of unidentified firms with 5-34 patents.

32 Matches from the second pass at unidentified firms (the 577/907 list).

41 "By-hand" matches to Compustat firms which are not in our current sample.

Not included in BYOTAF.D82 merge file

42 "By hand" matches to Compustat post-1976 acquired firms (41 cases).

50 "By hand" matches to non-Compustat firms (usually privately held).

51 "By hand" matches to foreign-owned U.S. firms.

52 Joint ventures (affiliate of at least 2 parent companies).

60 Investigated, but unmatched firms in the 577/907 list.

61 Located (address), but unmatched firms in the 577/907 list.

Finally, a list of OTAF organizations that are still not matched to Compustat firms was printed, along with the SRC code, if any. This file was sorted by decreasing number of total patents, and is named NOMATCH.C82. A quick glance at it shows that most of the top firms are foreign or U.S. Government (the U.S. Navy leads the list with 5934 total patents).

The top three firms with SRC=60 are RAYMOND LEE ORGANIZATION, NATIONAL RESEARCH DEVELOPMENT CORPORATION, and MARVIN GLASS & ASSOCIATES. These companies apparently obtain patents as a commercial service to individual inventors.

The top three firms with SRC=50 are HOFFMAN-LA ROCHE INC., HUGHES AIRCRAFT COMPANY, and LEVER BROTHERS COMPANY. These companies are privately held.

The output of this procedure was a tape (BYOTAF.D82 - AS2137) with a list of matched OTAF ID numbers and corresponding CUSIP numbers. We used this tape to select off a 750,000 record file from the Office of Technology Assessment and Forecasting which contained data on individual patents.

3.2 Updating the File in Summer 1982 (RNDPANEL.MASTER.AUG82)[4]

Work on the dataset in the summer of 1982 involved adding the patents data from the 1980 Patents Office tape, as well as adding additional variables from Compustat. The variables added are data items 48-59 (see the section "Dataset Description").

To add the patents data, it was necessary to select the observations from the patents tape that corresponded to one of the 5426 OTAF id numbers in BYOTAF.D82 (described above). Since the patents tape was in the form of one observation for each patent granted, it was necessary to aggregate the data to the OTAF level. By merging with BYOTAF.082, we switched from OTAF to CUSIP identification numbers.

The next step was to disaggregate to panel format: one record per CUSIP and year (1958-1980). For the years 1958 through 1964 and 1980, variables PATENTS and FORPAT (total patent applications and applications by foreign subsidiaries), and PATENTSG and FORPATG (total and foreign patents granted) were always set to missing. For the years 1965 through 1979 missing PATENTS and FORPAT, PATENTSG, AND FORPATG were set to 0. If the match of our sample to the OTAF tape is comprehensive, this coding for missing patents is correct, since the OTAF tape is complete for the years 1965 through about 1976, and partially complete from 1977 through 1979 (due to varying lengths of time from application date to date granted). However, because of the likelihood of our having missed some patenting by our firms, the zeroes in 1965 to 1979 must be taken as possibly indicative of missing

---

4. This section was prepared by Elizabeth Laderman.

patents rather than true zeroes, especially if the whole series is missing. In addition, the 1965 numbers actually represent the number of patents issued in 1966 through 1979 which were applied for in 1965 and earlier, so they are an overestimate of the number applied for in 1965.

3.3 Updating with 1983 Patents Data (RNDPANEL.MASTER.AUG83)[5]

We received a new patents tape in the spring of 1983, PATENTS.OTAF83.APR83 (AS0832), and used that data to update the patents information on the master tape. Following is a brief overview of the comparability of the old and new patent series.

The tape of matched old OTAF numbers and CUSIP numbers (BYOTAF.D82) contained 5426 observations; 3582 of these were matched to an OTAF number on the new patents tape by matching the old and new OTAF names. Of the 1844 unmatched old OTAF names, only 55 had an old patents total which was greater than two. Nineteen of these 55 were matched by hand to OTAFs on the new tape with almost exactly the same name. Thirty-three of the 55 were not worrisome because a new OTAF name that was very similar to the old unmatched name had in fact been matched. (For example, although Bliss + Laughlin Ind., Inc. found no match on the new tape, Bliss Laughlin Industries, Inc. did find a match.) In these cases it was assumed that the new series for the OTAF that was not found on the new tape was included in the new series for the OTAF that was found. The three major 1982 OTAFs that remained unmatched

---

5. This section was prepared by Elizabeth Laderman.

were General Time Corporation, E + B Incorporated and American Gage + Machine Company. So, counting the matches done by hand, 3601 out of 5426 observatins were matched at the OTAF level. Of the remaining 1825 observatinos, only 36 had old patent totals greater than two and only three of these 36 were troublesome.

After aggregating patents from the OTAF to the CUSIP level, 153 of the panel firms had positive old patent totals (when totaled over the 1965-1977 period) were missing new patent totals. None of these firms had an old patents total greater than four; the mean total was 1.568.

Excluding these firms, 77 (4%) of the 1904 CUSIPs on the master list had a change in total patents, from 65 to 77, that was greater than 20%. Of these 77, four had series sufficiently different that they deserved investigation. The others had so few patents that a difference of one or two patents in a few years resulted in a significant percentage difference in the totals. Two of the four firms were known troublemakers -- Katy Industries, the parent company of American Gage + Machine Company, and Talley Industries, the parent company of General Time Corporation. As expected, these two firms had many fewer patents in the new series. General Tire and Rubber Company and Zurn Industries had more patents in the new series than in the old series, for no obvious reason. The discrepancy for Zurn was not so great, but General Tire remains questionable.

In Table 3.1 the number of firms with patents, and the number of total patents for the old and new series are compared. This table was prepared after the updating in the next two sections was performed, so it also incorporates the lengthening of the sample to 1982 for most of the firms.

3.4 Adding the 1979-1981 Compustat Data (RNDPANEL.MASTER.JUL84)[6]

Work in the summer of 1984 consisted largely in adding data for 1979-81, from a new Compustat Industrial tape. In addition, we added some new variables for all firms and years.

We merged the 1959-78 and the 1979-81 data from the two Compustat files selecting only those firms which were on the old Compustat file. It is important to note that it was necessary to merge the raw data, because the SLEPIAN program requires a time series to calculate some data items: a continuous series could not be obtained by running SLEPIAN on the 1979-81 data and then merging. We then ran this data through the SLEPIAN program, slightly modified to account for the extra three years of data. Going back to the raw 1979-81 data, we added data items 50 through 57. The new file, in panel format by cusip and year for 1979-81, was merged with the patents data from the 1983 tape (described above). By merging this file with the old master file, we created an updated master with data for firms from the Compustat Industrial file through 1981.

---

6. This section was prepared by Joy Mundy.

TABLE 3.1

COMPARISON OF OLD AND NEW PATENT SERIES

	Total Firms with Patents			Total Number of Patents		
	Old	New	%	Old	New	%
1959	-	-		-	-	
1960	-	-		-	-	
1961	-	-		-	-	
1962	-	-		-	-	
1963	-	-		-	-	
1964	-	-		-	-	
1965	876	793	90.5	8,271	8,287	100
1966	923	834	90.4	11,984	11,987	100
1967	966	870	90.5	18,961	18,907	100
1968	1105	986	89.2	22,334	22,325	100
1969	1167	1033	88.5	23,285	23,300	100
1970	1218	1082	88.8	22,899	22,970	100
1971	1271	1124	88.4	22,243	22,302	100
1972	1354	1191	88.0	21,022	21,034	100
1973	1390	1223	88.0	21,280	21,306	100
1974	1431	1254	87.6	21,555	21,630	100
1975	1455	1273	87.5	21,099	21,345	101
1976	1494	1303	87.2	19,968	20,611	103
1977	1440	1303	90.5	17,418	20,393	117
1978	1362	1302	95.6	5,638	19,199	341
1979	350	1341	383	0	19,211	-
1980	-	1297	-	-	15,889	-
1981	-	1295	-	-	4,010	-
1982	-	0	-	-	0	-

3.5 Updating the File in December 1984 (RNDPANEL.MASTER.JAN85)[7]

The changes to the panel in December 1984 from two sources: IND - 30 firms that had moved out of the manufacturing sector on the 1981 tape, (and were therefore not picked up in the precious step) and OTC - picking up the years 1980-82 from the new 1982 tape.

The two sources of data were run through the Slepian sequence, separately. They were merged, and patent data added, and then concatenated to the old RNDPANEL.

We first identified suspicious firms by printing a list of firms on the '79 industrial tape that were not in the August 1984 panel, and those in the panel in 1976 but not in 1981. This list was checked by hand against Compustat's list of firms covered. It was discovered that some firms had moved out the manufacturing sector as of 1976, we wanted to pick up the 1979-81 data for these firms. We went through the same process as described in the previous section to add the 1979-81 data for these thirty firms.

The next task was to add the SPV deflators (price deflators for each firm and industry, data items 62-65) from a previously created dataset. These variables were added for years 1959-81. Fiscal year rates of return (RORF) were calculated from the quarterly Compustat files, and fiscal year-end price and cumulative adjustment factor (PRICEF and ADJFACTF) were also added for 1959-81.

Finally, we added some observations by including patents data from

---

7. This section was prepared by Joy Mundy.

1977-81 for all firms present in 1976. Some firms had patents data even after being dropped from Compustat, so these new observations contained only patents information (these observations were later deleted from the panel).

The process for adding the data for 1980-82 from the new OTC tape was identical to that described in the previous section for adding the 1979-81 from the Industrial file. We then merged these two distinct sets of new data and added the patents information. This set was concatenated with the August 1984 master and sorted by CUSIP and YEAR, to create the December 1984 master file. In January 1985 the final file (RNDPANEL.SAS.JAN85) was created by deleting about 500 stray observations which had only patents information and a few extraneous variables.

3.6 Adding the Compustat Data through 1985 (RNDPANEL.MASTER.OCT88)

In 1986-1987 a major updating of the R&D Master File of January 1985 was undertaken in order to create a continuous panel with data through 1985. Our original data sample definition was "publicly traded U.S. manufacturing firms which existed on the Compustat files for at least three years including 1976." This sample definition was changed to "publicly traded U.S. manufacturing firms which existed on the Compustat files (excluding the Full Coverage file) for at least three years sometime between 1976 and 1985." As mentioned before, omitting the Full Coverage file produces no serious loss of data. The effect of the change is to make this file a rolling panel of firms, with exits and entries every year between 1959 and 1985. See Table 1.2 for details on coverage.

Enlarging the sample was done in the following manner: since we had access to the Compustat Annual Industrial, Over-the-Counter, and Research files for 1981 through 1985 only, we drew the new sample from these files, including all manufacturing firms, but taking only the data from the most recent file on which the firm appeared. Lack of access to files from 1977 through 1980 weakens the coverage for very short-lived Over-the-Counter firms which do not end up on the Research files. Of course, our original sample was drawn from the 1978 Industrial file and 1980 OTC file, so the number of firms actually missing is probably very small.[8]

Originally, the primary reason for constructing this updated panel of firms was a study of mergers and R&D performance (see Hall (1988)). To find all the firms in this sample which were acquired during the period, it was necessary to study the exits from the sample; this allowed us to clean up the sample somewhat as a bonus. To find the exits, the headers (CUSIP, company name, year, etc.) from the new sample were merged with the old panel and all firms which exited from the sample before 1985 were listed along with the year of exit. All of these firms were looked up in various printed sources (see Hall (1988), Chapter 2, for details) and the reasons for exit found. The implications of this for the data construction project were the following: 1) some firms (about 90) changed names, but did not exit; we spliced these to their old records; 2) some firms did not actually exit, but dropped out of the manufacturing sector; we put them back in; 3)

---

8. For example, if we ask how many manufacturing firms of the 1294 on the original 1959-1978 Industrial File did not end up in our sample, the answer is only 4.

some firms (16) reorganized and changed their CUSIPs; again, we spliced these to their old records; 4) a few non-manufacturing firms (92) acquired manufacturing firms, and we wished to keep their data, so we went back to pick them up off the files. A separate file called MERGLIST.SAS.JAN87 was created at this point containing all the reasons for exit from the panel.

The new data (including all years for each firm) was run through the Slepian program (suitably updated) to create our constructed variables. This created a two files called ANNIND.SLEPIAN on tape with about 32,000 observations. These files were merged with the old panel in the following way: a few variables were renamed on the old panel for compatibility (DMERGE and CNAME), RNDEFLT was converted to 82 dollars, and NOSHARES was converted from thousands to millions. Then the new panel (ANNIND.SLEPIAN) was updated with the old panel data by CUSIP and YEAR. This has the effect of adding any observations and variables which were not on the new panel, but keeping any data on the new panel for which the old panel had missing values. This exercise yielded about 44,000 observations in a dataset called RNDPANEL.SAS.APR87.

The next step was adding the calendar year and fiscal year rates of return for these firms. Because these variables have to be computed using two adjacent years of data, it matters what file the observations come from, since Compustat may renormalize the data between two different data files owing to stock splits, etc. (The latest ADJFACT on a Compustat tape is always equal to one). Two files were created, each consisting of the firm's CUSIP, year of data, closing price of common stock, dividends, ADJFACT, and the computed rate of return. The first was RORC.SAS.NOV87, which was created from the merged and selected new

sample of 81-85, and the second was RORF.SAS.MAR87, which was created from the Quarterly Industrial Compustat files for 1981 and 1985. The creation of RORC took several steps, since it was necessary to merge in the data from the old panel and calculate the transition ROR extremely carefully (owing to the ADJFACT problem mentioned above). Because quarterly files are not available for the OTC file, we do not have fiscal year rates of return for these firms. RORC and RORF were merged with the panel to create RNDPANEL.SAS.NOV87.

In parallel with this last effort, a separate project was undertaken in order to construct a stock of R&D capital for these firms. Beginning with the March 1987 panel, the R&D series for each firm was extracted. Because of occasional missing data problems in the R&D series, which would cause all stock variables following to be missing, we interpolated the R&D series as described in the appendix. We required at least four observations on R&D to be present in order to form the interpolation, and the model we used was a random walk in the logarithms of real R&D expenditure, which was suggested by our previous work with this data. Once we had continuous R&D series, we constructed the end of period stock of R&D each year using a simple perpetual inventory declining balance formula with a depreciation rate of 15 percent.

Once RSTOCK was constructed, it was merged back into the panel in October 1988 to make RNDPANEL.SAS.OCT88. This file also contains a corrected and updated FILER code which gives the original Compustat source for each firm in the sample. In addition, the deflated R&D was rescaled back to 1972 dollars at this time, for consistency with earlier files.

REFERENCES

Brainard, W., J. Shoven, and L. Weiss. 1980. "The Financial Valuation of the Return to Capital." Brookings Papers on Economic Activity 2.

Directory of Corporate Affiliation. 1972, 1976. Skokie, IL: National Register Publishing Co.

Hall, Bronwyn H. 1988. "Research and Development Investment and the Evolution of the U.S. Manufacturing Sector: Econometric Studies at the Firm Level." Ph.D. diss., Stanford University.

Jaffe, S.A. 1972. "A Price Index for Deflation of Academic R&D Expenditures." NSF Bulletin 72-310.

Mergers and Acquisitions, The Journal of Corporate Venture. McLean, Virginia. Vol. 8-11.

National Science Foundation. 1978. Research and Development in Industry, NSF80-307. Washington, D.C.: GPO.

National Science Foundation. 1985. Science Indicators. Washington, D.C.: GPO.

San Miguel, Joseph. 1977. "The Reliability of R&D Data in Compustat and 10-K Reports." Accounting Review LII: 638-41.

Standard and Poor. 1978. Compustat. New York: Standard and Poor Corporation.

U.S. Dept. of Labor. Productivity Costs in Nonfinancial Corporations. Washington, D.C.: GPO.

Codebook for the RED Panel Dataset - RNDPANEL.DC188

VARIABLE	LABEL	N	MEAN	STANDARD DEVIATION	MINIMUM VALUE	MAXIMUM VALUE	STD ERROR OF MEAN	SUM
ACQUIS	Acquisitions (from stmt of changes)	1781	1.9371	18.725	-1.14600	495.116	0.44370	3450
ADJ	(CURR.LIAB.-DEBT)-(CURR.ASSTS-INVNTRIES)	2193	30.1403	163.387	-1490.50000	5184.324	3.48897	66098
ADJDEP	DEPRECIATION ADJUSTMENT	2193	33.5458	150.296	0.00000	3402.579	3.20943	73566
ADJFACT	ADJUSTMENT FACTOR FOR COMMON STOCK	2295	2.1813	2.415	0.00500	40.000	0.05041	5006
ADJFACTF	Adj. factor for common (fiscal year)	1249	1.5940	1.243	0.10000	13.200	0.03517	1991
ADJINV	INVENTORY VAL ADJ FOR INVENTORY METH(S)	2193	107.9691	348.712	0.00000	6505.070	7.44643	236776
ADJTOT	TOTAL DEFL. USING AGE STRUCTURE	2193	54.1650	247.340	0.00000	4033.152	5.28172	118784
ADPRICE1	IND PRICE DEFLATOR, ADJ FOR FYR	1904	1.4361	0.235	0.98362	2.570	0.00540	2734
ADPRICE2	FIRM PRICE DEFLATOR, ADJ FOR FYR	1904	1.4351	0.231	0.98362	2.591	0.00528	2732
ADV	ADVERTISING EXPENSE	1725	7.7238	28.662	0.00000	392.600	0.69011	13324
BKCAP	BOOK VALUE OF CAPITAL STOCK	2193	343.4506	1223.590	0.04300	24468.699	26.12064	753187
BKDEBT	BOOK VALUE OF LONG-TERM DEBT.	2295	86.2978	276.707	0.00000	3696.797	5.77603	198053
BKINV	BOOK VALUE OF INVENTORY.	2293	97.6795	324.416	0.00000	6327.801	6.77485	223979
BKPLNT	BOOK NET PLANT VALUE.	2295	194.2975	801.805	0.00000	18671.203	16.73699	445913
CASH	CASH AND SHORT-TERM INVESTMENTS	2292	43.1969	239.032	0.00000	6156.258	4.99285	99007
CIC	CUSIP ISSUE NUMBER AND CHECK DIGIT	2295	112.2898	44.190	0.00000	700.000	0.92243	257705
COMINC	INCOME AVAILABLE FOR COMMON	2293	29.3565	129.775	-46.34000	2809.901	2.71012	67314
COSTGOOD	COST OF GOODS SOLD	2291	448.0946	1813.934	0.00000	38164.309	37.89737	1026585
CURRASST	TOTAL CURRENT ASSETS	2283	231.2563	836.304	0.27500	15472.602	17.50295	527958
CUSIP	CUSIP NUMBER FOR FIRM	2295	499691.9198	298022.919	32.00000	989824.000	6220.97305	1146792956
DEPREC	DEPRECIAITON AND AMORTIZATION.	2292	20.8104	96.873	0.00000	2243.000	2.02347	47697
DFRTAX	Deferred taxes	1856	19.7431	85.004	0.00000	2149.929	1.97311	36643
DIV	DIVIDENDS PER SHARE	2288	0.4830	0.654	0.00000	8.000	0.01368	1105
DIVF	DIVIDENDS PER COMMON SHARE (FISCAL YEAR)	1249	0.6055	0.711	0.00000	8.000	0.02012	756
DLCOMP	LCOMP EXCLUDES EMPLOYEE BENEFITS	2295	0.0174	0.131	0.00000	1.000	0.00273	40
DLSEAS	Employees include seasonal figures	1882	0.0298	0.170	0.00000	1.000	0.00392	56
DMERGE	Major merger dummy (CS footnote)	2295	0.0266	0.164	0.00000	2.000	0.00341	61
DUP	CS code	1882	27.4368	43.173	0.00000	96.000	0.99519	51636
EMPLY	NUMBER OF EMPLOYEES IN 1000S	2237	9.3734	30.273	0.00800	748.000	0.64006	20968
EQUITY	COMMON EQUITY	2295	222.1676	853.997	-37.04700	18470.402	17.82645	509875
FILE	COMPUSTAT FILE CODE	2295	5.3913	3.355	1.00000	13.000	0.07003	12373
FORPAT	PATENT APPLICATIONS BY FOREIGN SUBS.	1494	0.3286	2.397	0.00000	47.000	0.06201	491
FORPATG	PATENTS GRANTED TO FOREIGN SUBS.	1494	0.3681	2.555	0.00000	54.000	0.06610	550
FYR	MONTH OF FISCAL YEAR END	2295	9.3107	3.451	0.00000	12.000	0.07203	21368
GRATE	GROSS RATE OF RETURN	2193	0.0875	0.082	-0.89683	1.239	0.00175	192
GROCAP	GROSS VALUE OF ADJUSTED CAPITAL STOCK	2193	731.1905	2752.594	0.19592	54434.645	58.77910	1603501
GROPLANT	GROSS PLANT.	2295	347.0067	1430.233	0.00000	29502.203	29.85489	796380
INCOME	INCOME BEFORE XTRY ITEMS	2293	29.6933	130.151	-46.34000	2902.801	2.71797	68087
INCTAX	Investment tax credit	1875	17.9130	83.973	-1.27600	1651.500	1.93928	33587
INVCAP	TOTAL INVESTED CAPITAL	2262	129.0994	710.941	-36.87198	10582.242	14.94815	292023
INVEST	CAPITAL EXPENDITURES.	2293	41.5980	188.517	0.00000	4098.391	3.93685	95384
LCOMP	LABOR COMPENSATION	374	453.5866	1090.350	1.20100	15286.602	56.38063	169641

Codebook for the RED Panel Dataset - RNDPANEL.OC188

VARIABLE	LABEL	N	MEAN	STANDARD DEVIATION	MINIMUM VALUE	MAXIMUM VALUE	STD ERROR OF MEAN	SUM
NETCAP	NET VALUE OF ADJUSTED CAPITAL STOCK	2193	477.7862	1741.144	0.06061	37112.563	37.18051	1047785
NOLCF	Net operating loss carry forward	1834	1.3781	10.235	0.00000	264.000	0.23900	2527
NOSHARES	COMMON SHARES OUTSTANDING	2289	9.6821	31.717	0.00400	706.118	0.66294	22162
NPLANT	DEFLATED NET PLANT VALUE.	2193	315.6523	1274.527	0.00000	29860.969	27.21634	692225
NRATE	NET RATE OF RETURN	2193	0.0536	0.098	-1.12073	1.368	0.00208	118
OLDPNL	This firm-year from 1981 Panel	2295	0.8296	0.376	0.00000	1.000	0.00785	1904
OPINC	OPERATING INCOME BEFORE DEPRECIATION	2292	81.4731	367.741	-21.90401	8698.902	7.68131	186736
PATENTS	PATENT APPLICATIONS	1494	13.8179	46.815	0.00000	831.000	1.21117	20644
PATENTSG	PATENTS GRANTED	1494	15.4712	50.855	0.00000	817.000	1.31571	23114
PENSION	PENSION AND RETIREMENT EXPENSE	1787	9.6172	41.610	0.00000	1071.200	0.98432	17186
PRICE	Price of common (calendar year)	2200	16.2101	17.416	0.06200	279.125	0.37132	35662
PRICEF	Price of common (fiscal year)	1238	19.4976	20.150	0.25000	279.125	0.57269	24138
PRICE1	IND PRICE DEFLATOR	1904	1.4307	0.231	0.98377	2.425	0.00530	2724
PRICE2	FIRM PRICE DEFLATOR	1904	1.4299	0.227	0.98377	2.591	0.00521	2722
RENTAL	RENTAL EXPENSE.	2090	7.7288	32.516	0.00000	613.943	0.71125	16153
RND	R & D EXPENDITURES IN CURRENT $.	1820	10.3003	53.179	0.00000	1257.120	1.24654	18747
RNDEFLT	DEFLATED R & D EXPENSE	1820	7.5682	39.074	0.00000	923.674	0.91590	13774
RNDFLAG	CODE FOR R & D CORRECTIONS	2295	0.0941	0.688	0.00000	6.000	0.01436	216
RORC	CALENDAR YEAR RATE OF RETURN	2128	0.4648	0.559	-0.94976	5.667	0.01212	989
RORF	Fiscal year Rate of Return	1218	0.3698	0.454	-0.77778	3.250	0.01302	450
RSTOCK	Stock of knowledge (RND) capital	1399	102.4250	461.709	0.01057	9953.280	12.34410	143293
SALES	SALES IN CURRENT $.	2293	605.1961	2326.407	0.00000	48631.113	48.58293	1387715
SHARESF	COMMON SHARES OUTSTANDING (FISCAL YEAR)	1249	14.3397	39.909	0.27900	706.118	1.12924	17910
SIC	COMPUSTAT 4-DIGIT INDUSTRY CODE	2295	3246.0915	942.105	1211.00000	9997.000	19.66564	7449780
STDEBT	DEBT PORTION OF CURRENT LIABILITIES.	2295	20.5293	94.439	0.00000	2108.234	1.97134	47115
STK	CS code, always zero	1882	0.0000	0.000	0.00000	0.000	0.00000	0
TL	LONG-TERM DEBT ADJ. FOR AGE STRUCTURE	2193	81.0078	256.384	-0.03832	3657.923	5.47485	177650
TOTAL	INVEST. IN UNCONS SUBS & OTHERS & INTANG	2193	39.1366	182.156	0.00000	3134.321	3.88977	85827
VAL	MARKET VALUE IN CURRENT $	2193	391.1308	1552.205	-2.10040	37226.516	33.14591	857750
VCOMS	VALUE OF COMMON STOCK IN CURRENT $	2206	311.4394	1441.425	0.00000	42062.684	30.68946	687035
VCOMS2	ALT. VALUE OF COMMON STOCK IN CURR. $	2295	231.2896	1275.021	0.00000	38560.445	26.61497	530810
VPS	VALUE OF PREFERRED STOCK IN CURR. $	2294	5.6676	36.142	0.00000	866.287	0.75459	13001
XREL	CS code	1882	764.1998	1795.112	0.00000	9999.000	41.37917	1438224
ZLIST	EXCHANGE LISTING AND S & P INDEX CODE	2295	4.6009	2.446	1.00000	20.000	0.05106	10559

APPENDIX 2

THE CONSTRUCTION OF THE R&D STOCK VARIABLE: INTERPOLATING THE MISSING VALUES OF R&D

To construct a variable that measures the stock of R&D capital owned by a firm, we use a method due to Griliches (1981; Griliches and Mairesse 1981, Griliches and Hall 1982). This method is based on a standard perpetual inventory equation with declining balance depreciation:

$$(1) \quad K_t = (1 - \delta) K_{t-1} + R_t$$

where K_t is the end-of-period stock of R&D capital and R_t is the (real) expenditures during the year. The depreciation rate δ is chosen to be fifteen percent per year; Griliches and Mairesse found that the exact choice of depreciation rate made little difference in production function estimates. This is not surprising since, if R&D expenditures are roughly constant in real terms, the stock of R&D capital is

$$K_t = \sum_{s=0}^{\infty} (1 - \delta)^s R_{t-s} = \delta^{-1} R$$

The variation across firms will then be approximately the same, regardless of the value of δ, and the magnitude of the coefficient will just vary inversely with δ. This means that separate identification of δ and the coefficient of K_t in an equation will be difficult.

Two missing data problems must be confronted when making a stock out of a series of flow variables: first, the problem of initial conditions for the stock, and second, the fact that a single missing value for R&D in one year will cause all the associated stock variables

to be missing. We solve the first problem by setting the initial stock to the R&D expenditures in the first year divided by the sum of the depreciation rate δ and a presample growth rate of new R&D of eight percent per year. Thus the individual stock is approximately four times the level of R&D in the first year. The second problem is solved as described below, by interpolation where there are only one or two missing values in an R&D series. This procedure affects relatively few firms.

The problem is that we may observe R_t and R_{t+s} where $s \geq 1$ but not $R_{t+1}, R_{t+2}, \ldots$. (R_t) is hypothesized to follow a random walk. How should we forecast $R_{t+1}, R_{t+2} \ldots R_{t+s-1}$, given values of R_t and R_{t+s}? The unbiased estimator is

$$E(R_{t+i} \mid R_t, R_{t+s}) = R_t + E\,(\varepsilon_{t+1} + \ldots + \varepsilon_{t+i} \mid (R_{t+s} - R_t))$$

$$= R_t + E(\varepsilon_{t+1} \mid \sum_{j=1}^{s} \varepsilon_{t+j}) + E(\varepsilon_{t+s} \mid \sum_{j=1}^{s} \varepsilon_{t+j}) + \ldots + E(\varepsilon_{t+i} \mid \sum_{j=1}^{s} \varepsilon_{t+j})$$

$$= R_t + \sum_{k=1}^{i} E(\varepsilon_{t+k} \mid \sum_{j=1}^{s} \varepsilon_{t+j})$$

where $\varepsilon_{t+k} \sim \text{IIN}\,(0,\sigma^2)$.

So we need to compute the conditional expectation of the disturbance ε_{t+k}, conditioned on a sum of s such disturbances, where the sum includes the $\{\varepsilon_{t+k}\}$ in which we are interested.

(NB: $E(\varepsilon_{t+k} \mid \sum_{j=1}^{s} \varepsilon_{t+j}) = 0$, where $k \notin [1,s]$).

To do this, note that

$$f(\varepsilon_{t+k}, \sum_{j=1} \varepsilon_{t+j}) = N\left\{\begin{bmatrix}0\end{bmatrix}, \begin{bmatrix}\sigma^2 & s\sigma^2\end{bmatrix}\right\}$$

so that

$$f\ (\varepsilon_{t+k} | \sum_{j=1}^{s} \varepsilon_{t+j}) = N(\sum_{j=1}^{s} \varepsilon_{t+j}/s,\ \sigma^2/s)$$

by the formula for the conditional bivariate normal, and hence

$$E(\varepsilon_{t+k} | \sum_{j=1}^{s} \varepsilon_{t+j}) = \sum_{j=1}^{s} \varepsilon_{t+j}/s.$$

This, in turn, implies that the optimal forecast of R_{t+i} is

$$R^*_{t+i} = R_t + (i/s)\ (R_{t+s} - R_t)$$

or $$R^*_{t+i} = R_{t+s} + ((s-i)/s)\ (R_t - R_{t+s}),$$

which is symmetric in the two endpoints as desired.

Now suppose $r_t = \log R_t$, the natural logarithm of R&D expenditures, and the r_t follows a random walk, i.e.,

$$r_t = r_{t-1} + \varepsilon_t \qquad \varepsilon_t \approx N(0,\sigma^2) \quad t = ..-2,1,0,1,2....$$

We know that

$$E(r_{t+i} \mid r_t, r_{t+s}) = r_t + (k/s)\ (r_{t+s} - r_t)$$

is the optimal forecast of r_{t+i}. What is the optimal forecast of R_{t+i}? The answer is that we must include the variance of ε_t in constructing such a forecast:

$$E(R_{t+i} \mid r_t, r_{t+s}) = \exp\left[r_t + i/s\ (r_{t+s} - r_t) + \sigma^2\right].$$

But this implies that we must have an estimate of σ^2, the variance of the shock, in order to construct the optimal interpolation of the series $\{R_t\}$. The simplest way to do this is to note that in general we have several observations on ε and that $\sigma^2 = E\varepsilon^2$, so that we can use the method of movements to estimate σ^2:

$$\hat{\sigma}^2 = \left[\sum_{t=1}^{t-1} e_t^2 + \sum_{t=t+s}^{T} e_t^2 + \left[s^{-1} \sum_{\tau=t}^{t+s-1} e_\tau\right]^2\right]/(T-1)$$

where the last term uses the information contained in the size of the jump over which we are trying to interpolate. σ_t denotes $r_t - r_{t-1}$, the data estimate of ε_t. Any other gaps in the data should be treated in the same manner. Given an estimate of $\hat{\sigma}^2$, we can form an unbiased forecast of R_{t+i} using all available data:

$$R_{t+i} = \exp\left[r_t + (i/s)\ (r_{t+s} - r_t + \hat{\sigma}^2)\right]$$

or

$$R_{t+i} = (R_t)^{1-(i/s)}\ R_{t+s}^{i/s} \cdot \exp(\hat{\sigma}^2(i/s)).$$

Once we have a continuous R_t series and an initial condition K_0, the stock series K_t is formed as in equation (2).