SMALL BUSINESS DATA BASE: PROGRESS AND POTENTIAL

Bruce A. Kirchhoff and David A. Hirschberg
Small Business Administration

SECTION I
INTRODUCTION

The Small Business Administration, Office of Advocacy, is responsible for carrying out a program of research and analysis to facilitate the growth of small business. PL 94-305, which created this office, directed the establishment of a program of economic research and analysis. Our charge involves understanding the impacts of governmental policy on small business growth and describing the condition of small business to governmental agencies and the Congress. Specifically, Sec. 202 of PL 94-305 lists the following functions:

1. Examine the role of small business in the U.S. economy, and the contribution which small business can make in improving competition, increasing economic mobility, restraining inflation, expanding employment, increasing productivity, stimulating innovation, promoting exports
2. Assess effectiveness of Federal assistance programs on small and minority business
3. Measure direct costs and other effects of regulation
4. Determine the impact of the tax structure on small business

More recently, Congress in PL 96-302 reaffirmed the required development of a small business "indicative data base" [Title I, Sec. 100 A(i)(5)], a small business external data base [Title IV, Sec. 401], and added a Presidential annual report to Congress on the status of small business [Title III, Sec. 303].

This latter requirement specifically asks the President to: (1) "present current and historical data," (2) "identify economic trends" and (3) "examine the effects on small business and competition of policies, programs and activities" of various government agencies.

This combination of legislative mandates clearly identifies objectives for the small business data base:

1. Provide a mailing list of small business in the United States.
2. Provide data describing the current condition of small business.
3. Provide descriptive data over time to identify trends.
4. Provide data for policy analysis.

Congress has not specified priorities among these objectives. It has specified in PL 96-302 detailed content of the four objectives itemizing both the economic and demographic data to be enumerated [Title IV] and the government policies it wants analyzed [Title III]. It leaves the source and form of the data base undefined but defines and budgets a separation of data into "indicative" and "external." And, the legislative history clearly specifies that no data collection burden shall be placed upon small business.

Federal Statistical Data

Congress has given us a unique assignment: "Build a data base without collecting any primary data." This is legitimate since our constituency is already burdened with redundant data collection paperwork imposed upon it by the Federal decentralized statistical system. Thus, we must obtain our data from other agencies. We describe below how we hope to do this.

Ideal "External" Micro Data Base

If confidentiality and other constraints were not factors, it is clear what we would obtain from the various Federal agencies for our micro data base. First, we would draw a sample from the Census Bureau's Standard Statistical Establishment List (SSEL) to accurately describe the small business population (probably about 250,000 firms). The SSEL is a comprehensive list describing the legal status and firm relationship of all establishments in the United States. Data would then be matched from the various five year economic censuses to obtain historical information on employment, sales, assets, and other components by firm. This data set would be matched with financial information from the IRS. For corporations the source would be the 1120 forms which, in addition to the tax liability information, provide key balance sheet and profit and loss information. Form 1065 provides partnership information; and data from the IRS 1040, Schedule C, F, and E provide information on proprietors (nonfarm and farm) and partners.

Next, to collect information on the characteristics of workers in these firms--sex, race, age, earnings, and work histories--Social Security data files would be merged. Finally, the

Federal Trade Commission's Quarterly Financial Report data would be merged to obtain current data on firms.

Such an exact match of records would provide researchers with a nearly complete data base as mandated by Congress. Also it would include micro data necessary for analysis of business response to various government tax, expenditure, regulatory and credit policies, as well as problems of inflation, recession, and productivity. However, present confidentiality rules of IRS, Census, SSA, and FTC make it impossible to obtain access to micro data.

Ideal "Indicative" Data Base

Access to the Standard Statistical Establishment List and to IRS Schedules C, E and F would solve our mailing list compilation efforts. These two files comprise a complete enumeration of the business population. Private data sources may approximate this but will never equal it. Most notably, the SSEL contains the enterprise/establishment linkages that are not complete or are not at all identified in other private mailing lists.

From Ideal to Real

Our ultimate goal is to use the Federal statistical system to develop the ideal data bases. It is cost efficient, avoids duplication, and eliminates any additional data collection paper work burden on business firms. There are however many barriers to our goal: (1) data aggregation, (2) restricted access, (3) different tabulation standards, (4) non-comparable reporting units in micro files, (5) time lags in data publication, and (6) incomplete data sets.

These issues are discussed briefly below:

Data Aggregation

The output of the Federal Statistical system consists almost entirely of aggregate and tabulated data. Although neat and tidy from a producer's point of view, it has an important shortcoming for policy analysis. It is not possible to determine if changes in distributions are due to behavioral changes or shifts in the mix of reporting units.

Access Problem

The lack of access to micro data has frustrated our development efforts since we first began. Virtually all agencies have statutes that restrict interagency transfers of micro data; i.e., data about any one individual business. Unless we can develop new legislation we will be left with a congressional mandate to obtain data from agencies that have congressional mandates to refuse our requests.

Federal Data Tabulation Standards

Each agency has adopted different employment, sales, and asset size standards, and these may change over time. Beginning in 1982, we have developed a definition of Statistical Business Size Categories for tabulating Federal statistics.

Noncomparability of Business Reporting Units

For the most part, each Federal agency that is charged with collecting statistics performs its work independently of other agencies. Industrial classification and geographic coding are generally not coordinated among the various statistical agencies. Thus collection methods differ in important ways that lead to major problems in combining data after collection.

Time Lag in Data Availability

Another difficulty in using existing Federal statistics is that availability is possible ony when they are published. This often occurs with a considerable time lag after collection.

Incomplete Data Sets

Last of all, most Federal statistical agencies collect data in accordance with their needs or Congressional authorization. "Size of independently owned business" is a relatively new concept that is not easily extracted from these existing data sets. Most notably, if we attempt to define size as number of employees, many sources of data are incomplete; i.e. without employment.

SECTION II
EXTERNAL DATA BASE - SBA INTERIM MICRO DATA FILES

Congress clearly sees the external data base as:

1. Providing data describing the current condition of small business
2. Providing descriptive data over time to identify trends
3. Providing data for policy analysis

Researchers would summarize these objectives as longitudinal (over time description) and cause/effect analysis. The Federal statistical system is widely respected for its ability to accurately describe business in the United States for over 40 years. But it is frequently criticized for lacking the data necessary to examine cause/effect relationships necessary for policy analysis. The President's Reorganization Project for the Federal Statistical System concluded that one of the major problems with the current statistical system was lack of policy relevance.

Congress has mandated that the small business external data base must be useful for policy analysis; therefore it must consist of micro data. Since micro data is easily aggregated by computer to

provide descriptive statistics, a micro data base maintained over time will adequately fulfill all three objectives.

Federal statistical micro data on business firms is simply not accessible to SBA under current confidentiality restrictions. Collecting our own data would be prohibitively expensive, e.g. the Bureau of the Census has a budget of $70 million for the 1982 Census of Business. Also Congress, in the legislative history of P.L. 96-302, clearly instructs Advocacy to avoid placing additional paperwork burden on business. Thus, we are compelled to use data collected by others.

This requirement of building a data base without actually collecting data represents a unique challenge. We have pursued several directions simultaneously and have gradually evolved these into three developmental categories: the SBA interim micro data base; Federal statistical system micro data; and special data development projects. These efforts are actually interrelated with each other and the Indicative Data Base. We will discuss each as a separate subject in this and the next section.

Design of the SBA Interim Micro Data Base

Congress has identified several uses of the micro data base. Most notable of these is preparation of data and analysis for the President's annual "Report on Small Business and Competition". In addition, a micro data base on business firms will undoubtedly be of interest to a wide variety of policy analysts. Once again, planning and system design requirements dictate that we identify probable users of this data.

The most important of these efforts involves the use of three Dun and Bradstreet files. These files are the cornerstone of the external and indicative data base effort. Dun and Bradstreet offer three separate data files as described below.

Dun's Market Identifier File

The Dun's Market Identifier (DMI) file contains information on business organizations that had financial activity in any one year. Each record in the file contains the following information on an establishment:

1. Dun's number - This is a number assigned by D&B that can be used to merge it with prior year's files.
2. Geographic location - City, county, state, SMSA, and zip code.
3. Year business started.
4. Annual sales volume.
5. Number of employees.
6. Standard Industrial Classification (SIC) and up to four minor SICs.
7. Parent and headquarter city and state.
8. Dun's number of parent and ultimate parent.
9. Subsidiary indicator.
10. Status indicator - Single location, headquarter, establishment, or branch.
11. Manufacturing indicator - Indicates whether or not manufacturing takes place at the location.

Dun's Trend Files

The Dun's Trend file consists of a set of variables for 600,000 firms appended to the DMI file. It includes for 1973 and 1978 the following variables:

1. Percent growth in sales.
2. Percent growth in employment.
3. Base year sales volume.
4. Base year employment.
5. Sales in 1978.
6. Employment in 1978.

Dun's Financial Statement Files

There are two Financial Statement files. The first consists of over 900,000 companies and provides data for one year. A longitudinal file is available for 324,000 companies containing data for at least two years. The variables in these files include:

1. Date of financial statement.
2. SIC numbers.
3. Number of employees.
4. Geographic location.
5. Year started.
6. Current and previous financial indicators (key balance sheet and profit data).
7. Cash.
8. Accounts receivable.
9. Inventory.
10. Notes receivable.
11. Current assets.

Dun's File Development

Dun's files present two important problems: First, the firms in the file are neither a census of all firms in the U.S. nor a random sample. Thus it is necessary to validate or "benchmark" the files against appropriate sources to be sure that the information drawn from the files accurately describes small business in total. Second, the files are not assembled by statistically rigorous data collection procedures, but instead by voluntary cooperation of respondents. Many firms provide incomplete data, and errors arise from a variety of sources. This makes the files "dirty"; some individual firm records contain missing or obviously incorrect data on one or more items. These records must be located, "cleaned," or rejected from the file before it can be validated or used for analysis.

Cleaning Dun's Files

Brookings Institution was contracted to perform this work. Their progress is detailed in a report entitled: "U.S. Establishment and Enterprise Microdata" (unpublished but copies available from SBA). Their work on the DMI file not only met the needs of the micro data base but also the indicative data base. In fact, for reasons dictated by computer processing and data consolidation, the indicative and external data bases are mixed and matched.

Brookings has successfully linked the three files to gain maximum information availability. They have developed a mechanism so establishments (places of business) can be identified with their appropriate enterprises (organizational units defined by ownership control). Basically this means that all of a firm's subsidiaries and branches (which are recorded as separate establishments in the DMI) are identified as belonging to the parent firm. This is necessary since it is size of parent that defines size of business. The indicative data base now contains a list of establishments defined by size of firm based on enterprise employment. Financial Statement file linkage to the DMI means that data on the DMI not on the Financial Statement file is now available and vice versa. In short, the linked files are far closer to meeting the objectives of a business micro data file.

There are still problems to be worked out. Some DMI employment figures were missing or inconsistent. Missing values have been imputed but inconsistencies remain. DMI sales are similarly afflicted. These problems must be studied and corrected.

Summary

For the reasons described, we are focusing our data base development efforts on the SBA Interim Micro Data files. Cleaning, validating, and extending these files longitudinally are now our major current activities. We hope to have a representative sample of 250,000 businesses in a clean, validated, partially longitudinal form ready for descriptive and policy analysis within FY 1982.

Still, the SBA-IMD will never be "finished." Every year new firms enter, old firms exit, and others grow or decline. This information must enter the longitudinal file as it accumulates current business activity in recognition that such activity will soon be history.

SECTION III
EXTERNAL DATA BASE: OTHER PROJECTS

In Section I, we described a multitude of problems associated with building a micro data base from Federal statistics. We are pursuing solutions to these and describe our actions below. Next we discuss data needs that are clearly necessary for policy analysis, but are not specifically identified in legislation.

Incomplete Data Sets

The preferred definition of business size is based upon total firm (enterprise) employment. However many data sets do not include this measure. At present three separate efforts are being made to add employment to existing data sets.

Imputing Employment into IRS Statistics of Income: A major limitation of IRS Statistics of Income (SOI) is that employment is not available. We intend to impute enterprise employment from one or another source onto the IRS micro data. If successful, IRS will retabulate their statistics by employment size, thereby increasing the descriptive information available.

FTC Quarterly Survey: Congress asked the FTC to reduce the paperwork burden it was placing on small business. In response, the FTC has reduced its sample size and simplified its form. As part of the form change, we have asked for collection of employment data. Questionnaires on these changes were sent out to small business leaders who showed no objection to the additional item. If employment is added, the QFR will be much more useful for examining sales, assets, and profits of small business. The FTC plans to ask for employment data beginning in October 1981.

Commercial Loans: Congress has asked the Federal Financial Institutions Examination Council to determine the feasibility of publicly describing depository institutions' commercial loan portfolios by size of business that is served. It is our hope that we can persuade these institutions to collect and routinely report information on the size of business (employment) borrower as part of the Federal statistical publications.

Self-Employment: 1960-75 Micro Data Samples

We have a longitudinal file on sole proprietors. This file is drawn from one made available by the Social Security Administration. Each year a one percent Continuous Work History Sample (CWHS), based on the same ending digits of the social security number, is drawn from individuals who file an IRS Form SE. This is a tax form for proprietors and partners who have earnings of more than $400 and have not paid the maximum social security tax from wage and salaried employment.

Included in each annual file is information on the sex, race, age, industry, county, and earnings of all covered proprietors. This longitudinal file is at the Bureau of Economic Analysis, Department of Commerce. Approximately 60,000 records are available each year. Because of recent interpretations by IRS of the 1976 Tax

Reform Act's confidentiality provisions data since 1975 have not been made available to CWHS users, including BEA. We would hope that this new confidentiality problem is resolved as soon as possible so that up-to-date information becomes available.

This data will allow description and trend analysis for policy purposes of a segment of small business that is not well described in any other Federal statistical program.

Longitudinal File of Workers by Size of Firm

Along with the Department of Labor's Bureau of Labor Statistics and Employment and Training Administration, an effort was completed to establish a longitudinal file of workers, also based on Social Security's 1 percent CWHS. This file contains a longitudinal quarterly earnings history for each job held by a sample of over one million workers. The variables included in the file are sex, race, age, industry, county, quarterly earnings. In addition special tabulations have been prepared which estimate the employment size of each business firm. This will show differences in the characteristics of workers in small and large companies for the first time.

Annual Survey of Manufactures

Richard and Nancy Ruggles have a two year grant to create a ten year longitudinal file of a sample of manufacturing firms in the U.S. Using Census of Manufactures and Survey of Manufactures data, they plan to build a file containing firm by firm micro data for each of ten years.

When complete the micro file created by the Ruggles will not be fully accessible by researchers. The file will be stored on a limited-access computer. Researchers will prepare analytical programs to examine the file, test these on a simulated sample from the file, and, when satisfied, submit these analytical programs to Census. Census will run the programs on the real file, review the results to assure no breach in confidentiality has occurred, and then give the results to the researcher. This form of limited access to micro data is the best Census can agree to under current confidentiality restrictions and is far greater than what is currently available.

IRS Proposed Access

To date no Federal agency has released business micro data, but our negotiations with IRS have led to a preliminary approach to see if sophisticated masking and sampling techniques can adequately ensure confidentiality, especially with small firms. For the publicly traded corporations, tax data are publicly available from the Security and Exchange Commission from 10K reports.

Special Data Development Projects

Congress was thorough in defining what it wants included in the indicative and external data base. However, other data are also required for adequate policy analysis. Thus we have initiated several special projects to develop policy-relevant data.

Summary Tabulation of History from the MIT Data Base

David Birch, Director of MIT's Program on Neighborhood and Regional Change, has worked with the DMI files for over six years. We have taken advantage of his expertise in several ways. Our first step was to request tabulations of base line data on the distribution of firms and establishments by size, by major industry, and by state and Federal region.

Gross Product Originating by Size of Business

This project for the first time provides annual industry estimates of Gross National Product for small and large business. The time series starts in 1955 and ends in 1976. Small business is defined as fewer than 500 employees, and medium and large business is defined as 500 or more employees.

Summary

As described in Section I, we are working towards an "ideal data base" built from Federal statistics. There are many barriers to be crossed as itemized in Section I. We have and are initiating projects to explore these barriers and crossing them one by one. In the meantime we find ourselves developing two data bases at once; our own indicative and SBA-IMD, and the ideal. We cannot meet the intent of P.L. 96-302 without pursuing these related but separate efforts. The cost of developing the ideal has thus far been over $500,000 per year, but as we proceed to surmount barriers, our opportunities for success grow.

*This is a summary of a more detailed paper by the authors available from SBA. Because of space limitations the mailing list project is not discussed.