THE DEVELOPMENT OF A BUSINESS SURVEY FRAME FROM ADMINISTRATIVE DATA

John Perry, UK Central Statistical Office
Room 1.015 Government Buildings, Cardiff Road, Newport, Gwent

KEY WORDS: Survey Frames, Administrative Data, Statistical Units

Summary

Administrative systems contain a wealth of information about the business community. This paper describes how a cost effective dynamically updated business survey frame (or register) for statistical purposes can be created using administrative data. It focuses on the specific example of the new register being introduced by the UK Central Statistical Office and Employment Department but also considers the wider issues in relation to developments in Europe. It considers how the differing needs of the administrator and the statistician for quality may be reconciled. The problems and advantages of open systems using relational database managers to deal with large volumes of data are explored. The paper also includes some consideration of the techniques needed to provide consistent selections for a wide range of statistical inquiries and to control the form-filling burden. Although the primary concern of the system must be to provide efficient selection, mailing and grossing facilities for statistical inquiries, the register is itself a valuable source of statistical analyses. The paper discusses the quality issues affecting such analyses, in particular lags in administrative systems.

Introduction

Creation of a business register for the conduct of statistical inquiries is an expensive process. Administrative systems contain a wealth of information about the business community. The resources available for the maintenance of such systems are generally far greater than can be afforded by a statistical office. Not only does the use of administrative data minimise the cost of the business register to the government agency but it also reduces the burden on businesses, not least in the creation and maintenance of the register.

The UK operates a decentralised system, whereby individual government departments have responsibility for their own statistical data collection. In the area of business statistics, there are three key data collection offices:

- the Central Statistical Office (CSO), which since 1989 has been responsible for the main inquiries supporting the national accounts,
- the Employment Department (ED) which has inquiries relating to labour market statistics, and
- the Department of the Environment (DOE), which collects information from the construction industry.

Each has its own requirements for business registers to support its statistical inquiries. Since the 1970's the Business Statistics Office (BSO), now part of the CSO, has used data from the Value Added Tax (VAT) system as a major register input. Prior to this it had relied heavily on the regional offices of the Department of Trade and Industry (DTI), for which the BSO was the main collecting agency for business statistics. Around 1980, the DOE decided that it could also use VAT for identification of construction businesses, when previously it had used its own regional office staff. The ED has for many years used information from the employee tax system ("Pay As You Earn" or PAYE) for the preliminary identification of employers.

During the 1980's the need to improve co-ordination of the data collection activities of the various statistical offices was recognised. One aspect of this has been the setting up of an Inter-Departmental Business Register (IDBR), which makes consistent use of administrative data and which provides a tool for the coordinated conduct of statistical inquiries. Approval to proceed with this project was given in 1992 and the first inquiries to use the resulting business register are due to commence in 1994 *(Economic Trends, April 1992)*.

Register requirements

The main purpose of the business register is the provision of a frame for sampling businesses for statistical inquiries. It must provide not only a sample but also information about non-sampled units

that allows inferences to be made about the business population. It must provide names and addresses for the despatch of forms (as business surveys are still conducted predominantly by mail). In addition the register can provide a good source of structural information in its own right.

To do its job the coverage of business activity must be known and duplication must be minimised. The register must be well-structured, with well-defined statistical units for sampling and for analysis. For efficient sample selection, industrial classification and a size indicator (preferably employment) are also needed. For despatch of inquiry forms, high quality mailing addresses are needed (in the UK this means that postcodes are included). Telephone numbers are desirable when it is necessary to contact businesses that have not responded to the statistical inquiry forms or when there are queries on returned forms. To control burden and to enforce non-response, information relating to selection and response is also needed.

Main administrative sources

The main purpose of the IDBR is the creation of a single source for the conduct of statistical inquiries within the CSO and ED. The DOE will also be a major user. A prototype system has been set up for use by the Northern Ireland Department of Economic Development (NIDED). The existing CSO business register (based on VAT) and the ED register (based on PAYE) will be replaced by the IDBR.

The use of separate business registers had resulted in inconsistencies in employment measures (and consequently in output per head estimates) and inefficiencies in maintenance. Part of the problem has been the use of different administrative inputs (VAT and PAYE). In the initial investigations, options were proposed that relied heavily on a single administrative source for the creation of the IDBR. Those sources considered were:

- VAT - traders registered for Value Added Tax
- PAYE - employers operating PAYE schemes
- Schedule D - businesses registered as self-employed for tax purposes
- Company Registrations - businesses operating with limited liability under the Companies Act.

As the existing source for the CSO business register, the quality of VAT (with 1.7 million registered traders) has been proved over several years. It has good although not universal coverage of business activity being weak in areas that are exempt from registration (mainly health and education services) and missing small businesses (with turnover below £37,600 in 1993/94) that register only voluntarily. The size indicator is turnover and industrial classification is reasonably good. New registrations and changes are supplied weekly (moving to daily within the next year) with no delay, once the trader has notified the local VAT office. Access to the data by the CSO for statistical purposes is allowed by the VAT Act 1983. The VAT number is used by the business in its trade with customers and suppliers.

PAYE (with 1.1 million registered employers) is a good source for counting employees but misses those businesses with low-paid or no employees, an important omission for the national accounts. The size indicator is employees (excluding the low-paid) and industrial classification although present is not an important part of the administrative system and as a consequence suffers in quality. Updating is quarterly and there are some, generally minor, delays in the notification of new businesses. Access to the data by the ED and CSO for statistical purposes is allowed by the Finance Act 1969. The PAYE number forms only part of the internal administration system.

The register of self-employed businesses (with around 3 million records) cannot be used by itself but could provide a supplementary source to PAYE. In theory, by combining the two sources, complete coverage of business activity would be achieved. In practice, there is little co-ordination between the self-employed and employer systems and there are legal constraints in the access to data on the self-employed by the statistical offices. The registration number is the personal national insurance number of the individual and thus does not relate directly to the business itself.

Registration of businesses is required under the Companies Act but only under certain conditions. Coverage in terms of business activity is good but small businesses, which tend to operate as sole proprietors or partnerships, are generally excluded from the registration process. Those that register are often inactive, with only about half of the 1.1 million registered businesses trading. There is no indicator

of size or industrial classification on the system, although this is changing with increasing automation. It is best thought of a supplement to other sources. The company name (generally with the suffix "limited" or "public limited company") is associated with the registration number and is available publicly.

The use of commercial sources was considered but rejected. Although they have the benefit of being publicly available, the lack of a statutory basis for their collection and maintenance means that both coverage and quality suffer in comparison with administrative data. Commercial sources have proved to be important, however, to supplement areas that are not covered adequately by administrative data. The primary use have been for maintaining structures of enterprise groups, for which Dun and Bradstreet provide a good source through their annual "Who Owns Whom" publications.

Various options were considered but, after discussion, the use of dual sources (VAT traders and PAYE employers, supplemented by a company number for incorporated businesses was chosen. Using either VAT or PAYE as a single source would result in unacceptably incomplete coverage either for the national accounts or for labour market statistics. The use of two sources creates problems of matching and consistency but it has one major benefit over a single source: it is not affected greatly by changes in the administrative source (for example an increase in the VAT registration threshold).

Counting the number of active businesses is not a simple task. The VAT trader source covers some 1.7 million businesses. PAYE provides information on around 400,000 additional businesses operating below the VAT turnover threshold or in areas that are exempt from VAT registration. Excluding the "black economy", we estimate that there are around 2.9 million active businesses in total. Those missing from the two main administrative sources but registered as self-employed include some workers within the construction industry, whose activity is picked up through statistical inquiries directed at the main construction contractors. Other businesses missing would in general be those operating on a part-time basis, sometimes in their spare time. The contribution to output is negligible. This has been checked on an occasional basis for the visible retail trades by the use of area sampling.

Dealing with consistency - statistical units

The administrative systems have their own business units that relate to the revenue requirements of the tax system and do not take heed of the statistician. After all, revenue must be maximised and operational costs minimised within the tax system. Within the VAT system, the individual legal unit (eg limited company, sole proprietor, partnership) is generally the tax unit. There are two exceptions: some companies register as groups, while a few register their operating divisions separately. The legal unit is well-defined but it is not necessarily the operating unit. As an example a publisher may set up separate companies for each imprint but on a day to day basis may treat publishing as a single business activity. Within the PAYE system, employers set up pay arrangements that do not necessarily relate directly to the legal unit. As in the example above, the publisher may employ all of the staff within one company within the group. Equally, there could be separate pay schemes for weekly and monthly paid workers.

It is vital for the consistency of the statistical inquiries to define business units that relate to the administrative units but which reflect the requirements of the statistical system. A consistent set of definitions for units also eases the maintenance task on the business register. The main units on the IDBR are the:

- *enterprise* - this has as its basis one or more legal units under common ownership, and
- *local unit* - this is the individual site (factory, shop, office etc) operated by the enterprise.

An enterprise must have at least one local unit.

The use of VAT as an input means that the enterprise has as its basis at least one VAT trader, where registered for VAT. It will often comprise one or more PAYE employer units. The administrative systems do not provide any information on local units and this must be obtained through the statistical data collection system. As most businesses operate at only one site, the burden on the business can be minimised by direct use of the administrative data to impute site details. The usual assumption is that enterprises with employment of ten or less operate at only one site. This is reasonably

robust for all except very fine small area statistics. The VAT address relates to the day to day operation of the business and can be used in general for the enterprise and, for enterprises with one local unit, as the site address.

In addition, the IDBR will support the *enterprise group*, which is defined as all legal units or enterprises under common ownership. For the conduct of inquiries, the enterprise group and even the enterprise may be too heterogeneous to be of use. Thus separate *reporting units* are set up to provide mailing addresses for forms and to define reporting structures where they form part of an enterprise. Most businesses are approached at the enterprise level, which is generally the natural level for reporting. Where part of an enterprise is required, the reporting unit defines the list of local units that represent the required coverage. Reporting units are generally used where homogeneity of reporting is required but may also be set up at the request of a business to ease its form-filling.

The use of standard units is consistent with the Regulation on statistical units of the European Community (*Official Journal No L76*). Although by itself the Regulation does not require Member States to use specific units, it sets down the definitions that form the basis of other legal instruments. A separate draft Regulation relating to the harmonization of business registers specifies the statistical units to be held on the register. The design of the IDBR is consistent with the proposals.

Dealing with consistency - size and classification

The administrative sources provide information on size and classification that can be used to reduce, if not eliminate, the burden on businesses, provided it is reliable. VAT provides an indicator of turnover that can be updated annually and a detailed classification based on a description provided by the individual trader or visits by the tax office. Both are a central part of the tax control system but even so are not completely reliable. Turnover, which is the value of outputs, is affected by outliers of two types: real atypical values and input errors. Quality checks of turnover are implemented to remove the worst errors and there is some additional feed-back, although after some delay, from the statistical inquiries. Industrial classification is more of a problem. Until 1993, traders were asked to enter their own industry codes (based on an out of date UK Standard Industrial Classification [SIC]). These VAT trade codes (VTC) were often inappropriate with many traders taking the easy way out and choosing a miscellaneous heading. This has been improved in the short term by the move to coding from descriptions by the local VAT office staff. Further improvements are expected in 1994, when the VAT classification system is due to be aligned with the latest (1992) revision to the SIC.

PAYE provides an estimate of employees based on a direct count from the administrative computer system. This is a recent introduction and previously employees were derived from tax estimates. The quality of the estimates has been the subject of detailed study, as they could provide direct estimates of employees without recourse to statistical surveys and thus reduce substantially the form-filling burden. Indications are that they provide an adequate size indicator for sampling, and hence as a register size indicator. Two problems have been identified, however: low-paid workers are excluded, even where they operate alongside other employees, and employees are retained on the tax records for some time after leaving the business. The industrial classification used by the PAYE system is based on the current SIC and is moving to the 1992 revision but the quality of coding is not high as it is not a requirement of the administrative system.

As far as possible, and despite the reservations about quality, the administrative data are to be used to maintain the register. Two over-riding considerations have led to this decision:

- the administrative data provide a consistent set of estimates, and
- the burden on businesses can be minimised.

As the administrative units are linked uniquely to enterprises, aggregation of employment and turnover from PAYE or VAT is possible. Industrial classification can also be derived from the administrative units, although in the case of PAYE this is not good. The limitation of the administrative sources means that it will be essential to conduct some selective register "proving" to check and supplement the information from VAT and PAYE.

The enterprise holds the following employment estimates:

- PAYE jobs - the sum of all employer units averaged over the four quarters to September
- Employment - PAYE employees adjusted on the basis of statistical surveys plus an estimate for working proprietors
- Employees - as returned or imputed for local units within statistical surveys

The employment estimate is the basis of sample selection and the other two estimates (PAYE jobs and employees) are used for quality and analytical purposes. The estimate of employment = constant * PAYE jobs + constant + working proprietors. Where the constants represent the relationship of PAYE jobs to employee estimates for local units responding to the large-scale annual employee survey. The constants vary by size, industry and location.

VAT turnover is the basis of enterprise turnover. It can suffer from double counting, where the sales from one company are purchases by another company within a group. Where this is the case or where it is not updated from the administrative source, survey responses are used. Industrial classification from VAT will be taken in preference to PAYE because of the quality considerations outlined above but statistical inquiries will be used more extensively to "prove" the coding. Where a business is registered for VAT but not for PAYE, employment is imputed from VAT turnover.

The effect of lags

Delays in the notification of changes to the register can have an adverse effect on the statistical inquiries conducted from it. Delays in the notification of births reduce the effective coverage of the register and can result in a downward bias in estimates of activity. Close liaison with those responsible for the administration of VAT means that the statistical register is effectively as up to date as the administrative system. Lags in the system can occur by late registration of businesses, although the severe penalties imposed on traders mean that the effect is limited. The VAT system requires, however, registration only once a trader has reached the threshold level. This means that small businesses may not be identified until they have been trading for some time. Deaths are more of a problem because deregistration for VAT purposes is not effected until the VAT problems have been resolved. It is possible through good response chasing within the statistical inquiries to deal with dead units still live on the register. Lags in changes to the size indicator (turnover) are not serious but can affect the efficiency of sampling. For births, 90% of registrations are on the register within six months. For deaths, the corresponding figure is nine months. Turnover is updated for more than 90% of traders within five months of the end of the tax year.

PAYE changes are notified to the statistical system only quarterly. Delays in new registrations are thought not to be serious, because employers tend to set up PAYE schemes in advance of employing staff. Deaths are a problem because schemes will continue even though there are no employees. The size indicator (employees) is updated dynamically on the basis of notifications from businesses. A change to zero could indicate a seasonal business or a potential closure of a PAYE scheme.

The use of the two administrative sources potentially reduces delays to notification of births. Adding the administrative units to the statistical system by imputing the statistical units prior to proving reduces lags. It can have the effect, however, of creating duplication of units and, through poor imputation, affecting the quality of sampling. Deaths can also be identified more promptly by the use of two inputs. Care must be taken as de-registration can be for purely administrative reasons. In addition the death of a PAYE unit does not necessarily indicate that the business has ceased trading, while a business de-registering from VAT could still have employees, even though it is trading below the threshold for VAT.

Statistical surveys from the register

The register provides a tool for producing consistency in inquiry results. It provides a consistent set of statistical units and size and classification for stratification. The CSO and ED conduct a very wide range of statistical inquiries, ranging from the monthly production sales inquiry (to support the index of industrial production) to the five yearly industrial purchases inquiry (for the input-output tables). Common reference numbers for the reporting and statistical units allow direct comparison of responses.

To support surveys with varying demands, the register is being maintained dynamically from the administrative sources, with extracts being taken where necessary for inquiry purposes. Inertia rules governing changes to turnover, employment and industrial classification operate to assist comparisons overtime.

The boundary between the register and the inquiry system has been defined. The register is designed:

- to provide analyses that assist sampling design,
- to provide the sampling schemes (based on stratification and simple random or probability proportional to size sampling),
- to select the units and associated statistical information for samples,
- to provide mailing addresses and despatch lists, and
- to link in to the forms printing system

Units selected for inquiries are marked as such and return of forms is indicated. The register can be used for despatch of reminders to non-responders and for enforcement. Control of overlap between consecutive despatches and between different inquiries can be achieved. Consistent mailing of forms is assured, through the administrative address, although flexibility to use other addresses is offered.

The register offers the ability to measure and control the form-filling burden on businesses. Overlap between inquiries to small businesses can be minimised and contact with larger businesses can be coordinated. A central feature of the monitoring of the performance of the IDBR will be analyses of form-filling.

Analysis from the register

The extensive use of administrative data to support the register means that the IDBR can be used to produce timely analyses of business activity. The existing CSO business register provides annual structural analyses of business activity *(Business Monitor PA 1003)* based strongly on data from VAT. It is proposed that this is expanded when the IDBR is operational to make fuller use of VAT and PAYE data.

The names and addresses of businesses themselves can be published only with consent through the statistical inquiries. As a result the CSO produces a directory of manufacturing local units (*Business Monitor PO 1007*) with scope limited by the coverage of the statistical inquiries and the willingness of businesses for names, addresses and industrial classification to be published. The IDBR will permit better coverage and identification of appropriate units for analytical purposes, but publication or release of information about individual businesses will still be constrained by the statutory framework.

The register can provide ad hoc analyses for many users within government, where their use of the information is within the legislative framework of the IDBR. For most users data may be used only for sampling for statistical inquiries or for statistical analyses but some strictly controlled access for administrative purposes is permitted. Supply of analyses to the European Community will be covered by a Regulation. Requests from other organisations can be satisfied if there is no disclosure, either directly or indirectly, of information about individual businesses. This limits the availability of the register and increases the cost of supplying analyses but it has been accepted that the main purpose of the IDBR is to serve the needs of central government for the national accounts.

Analyses of the VAT administrative data held on the business register will continue to be made available from work undertaken by the Small Firms Statistics Section of the DTI. These analyses (*CSO Bulletin 5/93*) have thrown light onto the life-span of businesses, although use of administrative data directly has some limitations. In particular, the notification of new businesses and of deaths of existing businesses is limited by the requirements of the administrative system. It is fair to state that most new registrations are received promptly, and sometimes in advance of actual trading. Deregistration, however, is a slower process, driven partly by the concern of the tax authorities to retain businesses as live while their tax affairs are being sorted out. Estimates for the effects of lags are made. A further problem is the question of what is a "real" birth. Administrative units may be deregistered and then re-register for reasons related only to the tax system. The enterprise will be a more stable unit but its stability will depend on the ability of clerical staff to identify what are real births. Any

analyses purporting to show changes in businesses will inevitably over-estimate the underlying change in the structure of businesses.

Information technology issues

Both the CSO and ED operate at more than one location and provide access to their business registers held on traditional main-frame computers remotely through a mixture of on-line and batch services. Input from the main administrative sources is currently through magnetic tapes. Developments in information technology in recent years have resulted in rapid increases in the capability of mid-range computers and communication systems to handle the volume of data involved in the maintenance of business registers. As a result of open tender arrangements, the IDBR will be installed on a Sequent computer running the INGRES relational database management system (RDBMS) under UNIX. Communications with users, generally operating with personal computers, will be through local and wide area networks.

Input from the VAT source will move from a weekly frequency on magnetic tape to daily using networks. The lower frequency but higher volume of some input, such as the quarterly update from PAYE, will continue in the medium term to be on magnetic tape as this remains a cost-effective approach. The use of the central data keying facilities will be limited, with most amendments coming either automatically from the administrative source or on-line from the end users.

Even with more powerful computers and relatively inexpensive disk storage, the inefficiency of RDBMS puts a strain on the system. A full size bench-mark test was set up to ensure the system would operate sufficiently quickly to process the large volumes of administrative data and numerous sampling runs. In addition a fully functioning prototype system covering data for Northern Ireland has been running for several months. The need to reconcile the differing demands of batch and on-line access have meant that batch runs will generally take place at night. On-line users can then expect the essential fast responses from the system during normal office hours. To prevent conflict with other applications, the IDBR will effectively have dedicated use of the computer. Resilience will be provided by cross-links between the register machine and other machines, which will come into effect automatically if any machine fails.

Communication between the machines within the CSO is ensured by compatibility of the computers and by an office-wide local area network. Easy communication by other (remote) users is essential and the necessary links have been set up at an early stage in the project.

Future developments

The project covers the existing needs of the CSO and ED for their statistical inquiries. Following major changes in the responsibilities of the CSO in 1989, many statistical inquiries have been taken on but not yet assimilated into the main business register. They are often small-scale and have their own register sources. As knowledge of the inquiries improves, it is expected that the sections responsible for them will move gradually to the IDBR. The ED is currently reviewing its strategy for employment statistics. As a result, inquiry needs will change over the next two to three years.

The IDBR development does not currently cover agricultural statistics, although the VAT source at least has good coverage of farm activity. There are no plans in the medium term for incorporating the farm registers held by the Ministry of Agriculture, Fisheries and Food (MAFF) but the position will be kept under review.

References

"The Inter-Departmental Business Register" - Economic Trends, April 1992 (HMSO)

Official Journal of the European Community No L76: Council Regulation (EEC) No 696/93

"Size Analyses of UK Businesses - 1992" - Business Monitor PA 1003, 1992 (HMSO)

"UK Directory of Manufacturing Businesses - 1993" - Business Monitor PO 1007, 1993 (HMSO)

"VAT Registrations and Deregistrations in the UK (1980-1991) - CSO Bulletin 5/93

PROBLEMS AND CHALLENGES ASSOCIATED WITH MANAGING A LARGE BUSINESS REGISTER

Paul S. Hanczaryk and Thomas L. Mesenbourg, Bureau of the Census
Thomas L. Mesenbourg, ECSD, Room 2584-3, Washington, DC 20233-6100

KEY WORDS: Business registers, organizational linkages, statistical units, Standard Statistical Establishment List, classification issues, business births and deaths

INTRODUCTION

The Standard Statistical Establishment List (SSEL) is a computerized list of all U.S. business firms and their establishments. It encompasses over 8.2 million establishment records and provides the basic frame for the economic censuses and over one hundred current Census Bureau surveys. The register is updated continuously with information from a variety of sources and serves a number of diverse uses and users.

The past 20 years has been an era of rapid change. New industries have emerged, older ones have disappeared. Many companies, faced with increased global competition, have fundamentally changed the way they operate as well as their organizational structure. These changes have resulted in a number of problems and challenges for statistical agencies managing business registers. While the SSEL has experienced many successes over the years, this paper will highlight some of the problems and challenges associated with managing a large register. The paper concludes with a brief description of some future improvement initiatives.

BACKGROUND

History

In 1968, the Office of Management and Budget designated the Census Bureau as the focal agency for the development and operation of a central business list, which we call the Standard Statistical Establishment List or the SSEL. This list of U.S. businesses, which became operational in the early 1970s, initially was intended to be shared with other Federal statistical agencies. Unfortunately, several attempts to enact legislation, which would provide access to the SSEL only for statistical purposes while protecting the confidentiality of individual businesses, have been unsuccessful.

Organizational Structure and Linkages

The SSEL maintains three organizational units and linkages:

Establishment--An establishment is defined as a single physical location within the United States where business is conducted or where industrial operations or services are performed. The SSEL contains records for approximately 6.2 million establishments with paid employees. In addition, we carry records for over 2 million establishments that have an Employer Identification Number (EIN) but do not have any current year employment or payroll. The establishment is the basic building block for the SSEL; establishments are linked to legal entities and to enterprises. The establishment unit permits us to publish statistics for detailed industries and small geographic areas.

Legal Entity--The legal entity is the organizational unit which, for tax reporting purposes, has been assigned an EIN by the Internal Revenue Service (IRS). Other government agencies, such as the Social Security Administration (SSA) and the Bureau of Labor Statistics (BLS), also use the EIN as a common identifier. The SSEL numbering system, which incorporates the EIN, allows us to link to, and draw from, the administrative systems of IRS, SSA, and BLS.

Enterprise--The entire economic unit consisting of one or more establishments under common ownership or control. The enterprise may be a single legal entity operating a single location or a complex family of legal entities and their establishments. Currently, the SSEL contains approximately 4.9 million single-establishment enterprises, and 165,000 multi- establishment enterprises that own and operate approximately 1.3 million individual establishments.

The SSEL does not contain units or linkages for subsidiaries (parent company owns 50 percent or more of the voting stock of another company or exerts control over the management and policies of the affiliate) or divisions (grouping of units defined by the enterprise; may or may not be a legal entity).

Information Contained on the SSEL

The SSEL includes over 100 program-relevant data items for each establishment which are regularly updated. Among the most widely-used data items are:

- Census File Number
- Employer Identification Number
- Company name, mailing address, and physical location
- Industry code, up to the 6-digit SIC code level
- State, county, and place geographic codes
- Legal form of organization code
- First quarter employment for last 3 years
- First quarter and annual payroll for last 3 years
- Sales/receipts data (census year and prior year only)
- Filing requirement codes

This represents only a partial list of the available data items maintained for each establishment. The SSEL also maintains similar information on legal entities and enterprises. Information required for processing (status files, hold files, MIS, and so on) is integrated into the register and is instantly accessible during processing. The SSEL as now structured uses about 75 gigabytes of disk storage.

OVERVIEW OF SSEL PROCESSING

The SSEL is updated regularly from administrative record information from the IRS, by information from SSA on new businesses filing for an EIN, by classification information from the BLS, and by a wealth of data collected in the Census Bureau current surveys and quinquennial economic censuses. In general, the information sources for the SSEL differ for multiunit (operating more than one establishment) firms and single-unit firms.

Single-Establishment Firms

Information for firms operating a single location or establishment is derived principally from the administrative records of other government agencies supplemented by Census Bureau current surveys. We obtain name, address, classification information, employment, payroll, and several other variables from the IRS under terms of our annual contracts. We obtain classification information for businesses filing for an EIN from SSA. In 1990, we began obtaining SIC codes from BLS for unclassified and partially classified EINs. The extensive use of administrative record data permits us to maintain the SSEL while minimizing the burden we impose on the business community. The Census Bureau's ongoing surveys, such as the Annual Survey of Manufactures and the Current Business Surveys, also provide considerable data for single-establishment firms.

In an economic census year, we collect extensive information from single-unit companies. The 1992 Economic Censuses will collect information from some 3 million single-establishment firms that exceed a predetermined annual payroll limit. In addition to the employment, payroll, and receipts data, the economic census mailings provide physical location information, complete SIC detail, and information on organizational changes. For example, the censuses reports will reveal single-establishment firms that now operate more than one physical location and thus are reclassified as multiestablishment firms.

Multiple-Establishment Firms

The SSEL obtains data on the current operations of firms operating more than one establishment (multiunits) essentially from the economic censuses and the annual Company Organization Survey (COS). All known multiple-establishment companies are mailed as part of the economic censuses to obtain employment, payroll, sales and receipts, industrial classifications, physical location information, company structure, and a host of other information.

In noncensus years, the Census Bureau conducts an annual COS to obtain basic data and to ensure that the organizational structure of each multiunit company is updated regularly. Typically, the COS mails report forms annually to all multiestablishment companies with over 50 employees, while companies with fewer than 50 employees are canvassed once between the censuses.

Administrative record data from IRS are also used to link to multiunit companies. Specifically, the EIN-level data from administrative record files are used to identify potential coverage problems as well as for editing.

PROBLEMS AND CHALLENGES

This section describes some of the most significant problems and challenges we encounter in managing, maintaining, updating, and improving our business register.

Maintaining Organizational Structure and Linkages

There are three major factors that cause difficulties in maintaining organizational structure and linkages on the SSEL:

- Substantial growth in the number of multiunit companies
- Increasing complexity of large companies
- Increasing demand for new organizational linkages to satisfy the demands of both data suppliers as well as surveys using the SSEL as a frame

Maintaining the organizational linkages among enterprise, legal entity, and establishment is the single most resource-intensive activity associated with the maintenance of the business register. The Census Bureau expends more than 80 percent of its SSEL resources maintaining the organizational linkages of companies operating multiple establishments.

Over the past two decades, the number of multiunit companies has grown tremendously. In the early 1970s, the SSEL contained 65,000 multiunit companies which owned about 540,000 establishments; by 1982, the number of companies grew to 129,000 with a million affiliated establishments. A decade later in 1992, the number of multiunit enterprises had increased to over 165,000 with 1.3 million owned establishments.

Not only have multiunit companies grown in number, the largest companies have become increasingly complex, exacerbating the problems of maintaining organizational linkages. The 1,000 largest, most complex companies, which own some 330,000 establishments, have undergone dramatic changes over the past decade. Mergers, acquisitions, divestitures, and wholesale reorganizations to meet increased global competition better have been the rule rather than the exception. The 1989 Recordkeeping Practices Survey indicated that decentralized management and decisionmaking made it more difficult for corporate headquarters to report knowledgeably about the activities of subsidiaries and their establishments, making data collection burdensome.

Finally, we have been faced with an increasing demand for new organizational units, both from data suppliers and internal Census Bureau users of the SSEL. For example, in the case of banking, insurance carriers, utilities, communications, and transmission pipelines, both the Recordkeeping Practices Survey and a 1990 Economic Census Pretest showed revenue and expense information was not available at the establishment level. In order to match census data requests with existing recordkeeping practices, we needed to develop new organizational linkages for these companies based on the legal entity operating within a state. In the 1992 censuses, we collected revenue and expense information from these new state-level units; we collect employment, payroll, and classification information at the establishment level. If the legal entity operated establishments in industries other than banking, insurance, communication, utilities, and pipelines, it was mailed a separate census form for each establishment.

In other instances, the inability to establish "special" organizational linkages easily has hampered some operations. For example, in the 1992 censuses some 150 companies took us up on our offer for special mailing arrangements, rather than mailing all forms to the parent company. The capability of linking selected establishment records with these special units would have saved us countless hours of programming and manual labor associated with the assembly of special mailing packages.

Finally, over the past several years there has been an increasing demand for subcompany organizational units to meet the needs of those current surveys that collect financial information. Surveys such as the Quarterly Financial Report (QFR), the Annual Capital Expenditures Survey (ACES), and others would benefit if subcompany organizational linkages were available on the register.

Identifying Organizational Changes--Business Births and Deaths

To date, we have not been able to provide reliable statistics on business births and deaths from the SSEL. For multiunit companies, the annual Company Organization Survey and the Economic Censuses provide comprehensive information on organizational changes and track these changes through a Permanent Plant Number, which indicates the initial identification number of the establishment.

We have not yet developed a similar mechanism to track single-establishment records over time. Although we can identify new activations from administrative record sources (as indicated by new EINs), these new units do not necessarily represent true births. Many represent legal reorganizations.

As an example, many sole proprietorships and partnerships eventually become corporations, which requires a new legal entity represented by a unique EIN. We would like to develop a methodology to link these and other types of business reorganizations across years --an accurate method to distinguish true births from reorganizations is essential in providing meaningful birth/death statistics.

Here again, funds are limited, but we have begun to evaluate the feasibility of developing a linked file that could be used to produce birth/death statistics. Specifically, we are evaluating statistical matching techniques to provide the necessary linkages for developing comprehensive statistics on business births and deaths.

Classification Issues

Probably one of the most complex problems associated with the SSEL is the assignment and maintenance of accurate industry codes for all organizational units. Classification codes on the SSEL are derived primarily from six sources:

1. Economic Censuses--Census questionnaires are mailed to all establishments of multiunit companies and to approximately 3 million of the 4.9 million single-establishment companies with current year payroll. In general, codes are assigned based on detailed product, commodity, merchandise line information, or kind-of-business information.

2. Census Bureau Current Surveys--Generally, SIC classification codes are assigned based on written descriptions.

3. Social Security Administration--SSA assigns SIC codes to businesses applying for an EIN based on written description of principal business activity. SSA codes approximately 900,000 EINs a year. Codes are available to the Census Bureau about 6 months after the initial EIN application.

4. Bureau of Labor Statistics--The Business Establishment List (BEL) maintains and updates SIC codes for about 5 million employer establishments. BLS codes about 500,000 new establishments every year, and annually about one-third of existing establishments are reviewed and SIC codes are updated. Under a special arrangement, BLS provides Census quarterly with SIC codes for unclassified or partially classified manufacturing SSEL establishments.

5. Internal Revenue Service--Business entities (corporations, partnerships, and sole proprietorships) select a code from a list of principal business activities.

   These codes are less detailed than the SIC industry codes. For example, there are 1,005 SIC industries, 187 corporation codes, 199 partnership codes, and 184 sole proprietorship codes. IRS codes for corporations and partnerships are provided to Census annually, sole proprietorships only quinquennially. Before being carried to the SSEL, all IRS codes are converted to a SIC basis.

6. Business Name Coding--An automated name-coding program assigns SIC codes to SSEL records based on certain key words and phrases in the business name. Those cases not coded in the automated program are clerically coded.

   In terms of assigning and maintaining classification codes on the SSEL, we face a number of problems:

   - Updating classification codes
   - Increasing number of unclassified units
   - Classifying legal entities, subsidiaries, and enterprises

Updating Classification Codes

Clearly, one of the challenges we face is integrating classification codes which are derived from various sources into the SSEL. While all four agencies use coding systems based on the SIC, it is not surprising that the same unit may be coded differently by different agencies. Differences in the source documents used to assign codes, in coding procedures, differing resources and expertise, and different program objectives explain many of these discrepancies. Matching studies and evaluation of codes from differing sources have explained some of these differences, but more research is clearly needed. Currently, if we have codes from different sources, we use the hierarchy described above, with the codes from the economic censuses being given priority. Research from the 1992 Economic Censuses will be used to reassess this hierarchy.

Increasing Number of Unclassifieds

Over the past 5 years, we have seen a rapid increase in the number of unclassified businesses on the SSEL. For example, in 1987 the number of unclassified establishments on the SSEL was approximately 170,000. By 1989, the unclassifieds had increased to 400,000; and by 1991, the number had grown to over 650,000. The number of unclassifieds always increases after the census because we do not obtain sole proprietorship codes annually. Recent changes in procedures for providing new EINs and assigning SIC codes also contributed to the increase.

We have effectively reduced the number of unclassifieds using a three-pronged approach. Beginning with data year 1990, we provide BLS quarterly with a file of our unclassified EINs as well as some partially classified records. BLS matches these records to its BEL and returns codes back to the Census Bureau for about two-thirds of the cases. The remaining unclassified records are run through an automated batch program that assigns SICs using the business name. This routine assigns codes to about 25 percent of the unclassified records. Cases that are not coded in the automated program are then clerically coded through an interactive SIC routine. These processes typically reduce the unclassifieds from 650,000 to about 125,000. In an economic census year, we reduce the unclassified component further by mailing classification cards to the remaining uncoded cases.

Classifying Legal Entities, Subsidiaries, and Enterprises

Approximately 80 percent of the SSEL universe consists of single-establishment companies with a single EIN. For these companies, the enterprise, the legal entity, and the establishment are identical. Companies mailed in the censuses generally have 4-digit SIC codes, while nonmail cases will have less detailed codes based on administrative record sources.

Multiunit companies may consist of one or more legal entities which are represented by EINs. The 165,000 multiunit companies currently have over 300,000 active EINs. Our original plan, not yet implemented on the register, was to code multiunit legal entities by summing payroll for all the individual establishments within the EIN with the same 4-digit SIC code, compare industry totals by 4-digit codes, and assign the EIN the 4-digit SIC code of the industry with the largest total payroll. This methodology would be useful in constructing frames from the SSEL where the legal entity is the appropriate statistical unit. For the largest, most complex companies, a secondary activity code also would be computed. The largest 1100 companies have over 32,000 EINs. Nearly 55 percent of those EINs include establishments all operating in the same 3-digit SIC; however, 20 percent of the EINs include establishments operating in five or more distinct industries.

Currently, the SSEL does not maintain an organizational unit for the subsidiary or division. As we develop these new organizational units, a prerequisite will be that these units be defined in terms of establishments. In this way, as establishments are updated by various programs, we will be able to update the SIC code of the reporting unit.

The enterprise code uses the establishment SIC code as the basic building block. Currently, enterprise codes are assigned on an "as needed" basis and do not reside in the database. There are two primary ways we assign enterprise codes, the "filter-down" or "filter-up" methods. The most common method used by the Census Bureau is the "filter-down" process. Under this method, we sum individual establishment data, generally payroll but sometimes employment or sales, to ascertain the largest industry division (manufacturing, retail, and so on). Once the largest division is determined, individual establishments are summed to the appropriate enterprise category, and the largest category is assigned to the enterprise. The enterprise code is always derived from the largest division.

Under the "filter-up" method, each establishment is converted to an enterprise category code, identical categories are summed, and the largest category is assigned to the enterprise. The enterprise code for the company is not necessarily from the largest division.

Even when the same method is used to assign enterprise codes, the results may differ significantly if different variables are used in the determination. In fact, one of the biggest challenges we face in improving the usefulness of the SSEL as a company and subcompany frame is related to this point. For example, the QFR classifies corporations using the "filter down" method but assigns SIC codes based on gross receipts rather than payroll. Annual receipts information is not currently available on the SSEL except in census years and the prior year. To serve the needs of the QFR and several other current surveys, we need to add receipts to the SSEL on an annual basis.

RESOURCES

The SSEL was initially funded in the early 1970s. Since that time, the frame has grown significantly in terms of the number of records, complexity, and the demands placed upon it. One recent example was the additional requirement associated with the expansion in economic census scope to include coverage of finance, real estate, insurance, communications, utilities, and those transportation industries not covered in the 1987 censuses. Even though the workload, complexity, and demands on the SSEL have grown significantly over the past two decades, funding has not increased accordingly. In fact, although several requests for additional funding have been submitted, no supplemental funding has ever been provided. The only changes in the base budget reflect adjustments for pay raises and inflation.

Faced with the prospect of relatively flat funding, we concentrated on automating much of the processing associated with register maintenance. By automating processes which were manually-intensive, time-consuming, and complicated, we freed resources for further improvements in the SSEL. Beginning with the 1982 Economic Censuses, we moved aggressively from a forms- and paper-based, batch-oriented processing environment to a database environment, highly dependent on interactive access. Faced with programming and subject-matter constraints, our objective was to develop a processing system that would support both SSEL and economic census processing needs. Personnel savings realized in our processing operations were reinvested in programmers, hardware, software, and other critical resources.

During the 1987 Agriculture and Economic Censuses, our processing environment began to change again as we moved from a UNISYS platform and began using new DEC minicomputers and new database tools. In 1989, we began the migration of the SSEL and the associated processing systems to DEC. The DEC environment provided new tools and improved methods of processing, but also required substantial training and the conversion of all existing interactive routines as well as the development of a number of new ones. For the 1992 censuses, over 40 interactive routines have been implemented on DEC, and a number of other processing subsystems have also been converted including geocoding, data entry, and disclosure.

What, one might ask, is the problem? The problem is, of course, the SSEL as it currently exists is inextricably linked and enmeshed with a complex, ever-evolving processing system constructed to update and maintain the register. In a census year, the complexity of the processing environment increases immensely. Census mailout, checking in report forms, remailing delinquent businesses, data capture, updating the SSEL, releasing data to subject divisions, and so on, complicate the processing environment tremendously in contrast to a noncensus year. Consequently, significant increases in processing requirements place such heavy demands on our programming and subject-matter staffs that current survey requirements often cannot be satisfied in a timely fashion. For example, the demands of ensuring that the census follow-up programs were tested and in place, of bringing the telephone system online, and so on, have so fully occupied our programming staff that we had to defer a number of current survey requests by several months.

Ideally, we would like to separate the SSEL as a frame from the control file processing with which it is so inextricably linked. The SSEL as frame would stand independent, while a mirror image of the SSEL would be part of the census control file processing system. The maintenance and updating of the SSEL would be identical in a census and a noncensus year. The decreasing cost of storage and the availability of the necessary computer hardware make this vision a possibility. All we need is the additional resources to make this a reality.

FUTURE OPPORTUNITIES

Clearly our plate is full. However, we are hopeful that we cannot only overcome some of the current problems and challenges we face but also can improve the usefulness, quality, and maintenance of the SSEL. Space does not permit a detailed explanation of all of our current SSEL initiatives, but we mention a few of the most important below.

- Restructure the SSEL--After completion of the 1992 Economic Censuses, current plans are to restructure the SSEL. Under the proposed structure, we will implement new organizational linkages for multiunit companies. Evaluation work is underway using the IRS information on parent companies and their affiliated corporations. We also will implement the capability of developing special linkages between subsets of establishments to handle unique mailings, to facilitate electronic reporting, or satisfy other demands.

A FY 1995 budget initiative requests funds for expanding the content and coverage of the SSEL. This initiative would expand coverage of the SSEL to small businesses with no paid employees and add annual sales/receipts data to the SSEL. The new structure will include these additions.

- Another priority is to separate SSEL maintenance and updating from the processing subsystems that now are so inextricably linked to the register. Another way of approaching this goal is to think of it as separating the uses of the SSEL from the processing subsystems. For example, we would develop improved interfaces that could access the SSEL and satisfy a variety of purposes. The first organizational steps necessary to accomplish this have been laid out, but full implementation requires additional resources.

- Classification Improvements--The United States has undertaken a major initiative to restructure the industry classification system for 1997. We are hopeful that the new system will mirror the present structure of the economy better, answer much of the criticism of the current SIC, and meet the needs of 21st Century data users. We plan to implement multiple classification code fields on the register for each establishment to meet myriad analytical requirements better. Thiscapability also will be available for all other organizational units.

- Data Sharing--New and more effective ways of sharing data among agencies for statistical purposes, while protecting confidentiality, must be actively pursued. The obstacles are significant, but the opportunities and potential benefits demand that we address this issue again.

- Total Quality Management (TQM)--We are committed to continuing our TQM journey. We are working with our key suppliers--the IRS, the SSA, and American businesses--to improve long-standing relationships. We have begun developing a systematic approach towards measuring, monitoring, and continuously improving the quality of the register. Finally, we will continue to work closely with our customers, both inside and outside the Census Bureau, to ensure that the SSEL meets their needs.

USE OF CAPTURE-RECAPTURE TECHNIQUES TO ESTIMATE POPULATION SIZE AND POPULATION TOTALS WHEN A COMPLETE FRAME IS UNAVAILABLE

Kenneth H. Pollock, North Carolina State University
Steven C. Turner and Craig A. Brown, National Marine Fisheries Service
Kenneth H. Pollock, Department of Statistics, North Carolina State University, Box 8203, Raleigh, NC 27695-8203 USA

KEY WORDS: Incomplete Frames, Capture-Recapture Sampling, Large Pelagic Survey, Angler Surveys, Telephone Surveys, Access Surveys.

ABSTRACT

In classical sampling theory it is assumed that a complete frame exists. That is there is, at least conceptually, a complete list of population units. It is then possible to draw a probability sample from the population. Estimators of population parameters such as mean or total then have known properties and are easily studied theoretically or numerically.

In practice in surveys of establishments a complete frame may not exist. Lists of establishments kept by professional associations or government agencies are often incomplete. One approach to tackling this problem is to use the multi frame approach originally developed by Hartley (1962, 1974). An example of this approach is the National Agriculture Statistics (USDA) farm surveys. These surveys use an incomplete list frame of farms plus an area frame where all farms within a sampled unit are enumerated. Therefore the list frame is incomplete while the area frame is conceptually complete because there is a list of all area units and within each area unit theoretically all farms could be enumerated.

There are some situations, however, where it may not be possible to use an area frame for practical reasons. All that the researcher may have available may be several incomplete frames of establishments. The usual approach in this situation is to merge all the incomplete lists and ignore any remaining incompleteness. Depending on the degree of incompleteness remaining there could be serious negative bias on estimates of population size and population total.

In this manuscript I present a formal model based sampling solution to this problem based on capture-recapture sampling. Capture-recapture sampling models are widely used in sampling animal populations and also for adjusting the US census for undercoverage. In the simplest case of two incomplete lists we consider "marked" units to be those which occur on both lists and unmarked units to be those which do not occur on both lists. It is easy to estimate total frame size using the Lincoln-Petersen estimator. This estimator is model based with a key assumption being independence of the two lists.

Once an estimator of the population size has been obtained it is possible to obtain an estimator of a population total for some characteristic if a sample of units have that characteristic measured. The usual estimator of a population total for simple random sampling without replacement is $N\bar{y}$ where N is known and $\bar{y}$ is the mean of the sample. Here our estimator is $\hat{N}\bar{y}$ where $\hat{N}$ is obtained from the capture-recapture approach. This means that bias and precision of the estimator may be due to both components ($\hat{N}$ and $\bar{y}$). A discussion of the properties of this estimator will be presented.

An example of where this approach has been applied is a National Marine Fisheries Service Large Pelagics Survey. In this survey the establishments are fishing boats taking part in the ocean fishery off the Atlantic Coast of the United States. Several incomplete lists of boats are used to form the capture-recapture estimate of population size. Population totals for numbers of fishing trips (effort) and number (or weight) of fish caught (catch) are of primary interest.

Estimation of frame size and then population totals using a capture-recapture model is likely to have broad application in establishment surveys. The advantages are obviously practicality and cost saving. The disadvantages are obviously possible biases due to assumption violations. Our philosophy of using a model based approach to estimating a non sampling error is not new and is now being widely applied to studying many other non sampling error problems.

1. INTRODUCTION

Despite the assumption of the existence of a complete frame by most sampling theory textbooks (Cochran 1978) there are many real surveys (including those of establishments) where a complete frame does not exist. In the next section we consider classical sampling theory and incomplete frames. We suggest the possibility of using capture-recapture methods to estimate frame size. In Section 3 we review the capture-recapture literature to give an overview of the types of models available. In Section 4 we present an example of a sample survey of fishing boats. (We consider a boat analogous to a business establishment). While this example has some unique features we believe it has many features common to other establishment surveys. In the final discussion section we summarize the strengths and weaknesses of using the capture-recapture approach to estimating frame size in establishment surveys. Many of our ideas will require further research.

2. CLASSICAL SAMPLING THEORY AND INCOMPLETE FRAMES

In classical sampling theory it is assumed that a complete frame exists. There is, at least conceptually, a complete list of population units. It is then possible to draw a probability sample from the population. Estimators of population parameters such as mean or total then have known properties and are easily studied theoretically or numerically. Books on sampling theory such as Cochran (1978) concentrate on this situation and give properties of estimators for common sampling designs such as simple random sampling, stratified random sampling and multi-stage (cluster) sampling.

In practice in surveys of establishments or businesses (such as fishing boats) a complete frame may not exist. Lists of establishments kept by professional associations or government agencies are often incomplete. One approach to tackling this problem is to use the multi frame approach originally developed by Hartley (1962, 1974). An example of this approach is the National Agriculture Statistics (USDA) farm surveys (references). These surveys use an incomplete list frame of farms plus an area frame where all farms within a sample unit are enumerated. Therefore the list frame is incomplete while the area frame is conceptually complete because there is a list of all area units within each area unit theoretically all farms could be enumerated.

There are some situations, however, where it may not be possible to use an area frame for practical reasons. All that the researcher may have available may be several incomplete list frames of establishments. The usual approach in this situation is to merge all the incomplete lists and ignore any remaining incompleteness. Depending on the degree of incompleteness remaining there could be serious negative bias on estimates of population size and population total. (This is certainly true for the fishing boat owner telephone-access survey discussed later).

Later we present a formal model based sampling solution to this problem based on capture-recapture sampling. Capture-recapture sampling models are widely used in sampling animal populations (Seber 1982) and also for adjusting the U.S. census for under coverage (Wolter 1986). In the simplest case of two incomplete lists we consider "marked" units to be those which occur on both lists and unmarked units to be those which do not occur on both lists. It is easy to estimate total frame size using the Lincoln-Petersen estimator (Seber 1982, p. 59). This estimator is model based with a key assumption being independence of the two lists. Once an estimator of the population size has been obtained it is possible to obtain an estimator of population total for some characteristic if a sample of units have that characteristic measured.

The usual estimator of a population total for simple random sampling without replacement is

$$\hat{Y} = N\bar{y} \qquad (2.1)$$

See for example Cochran (1978, p. 21) where N is known and $\bar{y}$ is the mean of the sample. The variance of $\hat{Y}$ is given by

$$\mathrm{Var}(\hat{Y}) = N^2\mathrm{Var}(\bar{y}), \qquad (2.2)$$

where $$\mathrm{Var}(\bar{y}) = \frac{S^2}{n}\left(\frac{N-n}{N}\right).$$

and S^2 is the population variance and $\left(\frac{N-n}{N}\right)$ is called the finite population correction factor. The estimator (2.1) is also an unbiased estimator of the population total.

Here our estimator is

$$\hat{Y} = \hat{N}\bar{y} \qquad (2.3)$$

where $\hat{N}$ is obtained from the capture-recapture method.

This means the properties of the estimator (2.3) is more difficult to evaluate because both $\hat{N}$ and $\bar{y}$ are random variables unlike

in estimator (2.1) where N is a known quantity. The estimated variance of $\hat{Y}$ here is given by

$$\hat{Var}(\hat{Y}) = (\hat{N})^2 \hat{Var}(\bar{y}) + (\bar{y})^2 \hat{Var}(\hat{N}) + \hat{Var}(\bar{y}) \hat{Var}(\hat{N}) \quad (2.4)$$

assuming that $\bar{y}$ and $\hat{N}$ are independent and using a result due to Goodman (1960). The estimator (2.3) is only an unbiased estimator if $\hat{N}$ and $\bar{y}$ are unbiased estimators of the population size and population mean respectively which is not usually the case in practice. We discuss the estimator (2.3) and some generalizations further when we discuss the use of the capture-recapture method in the large pelagic fishery survey example.

Estimation of frame size and then population totals using a capture-recapture model is likely to have broader application in establishment or business surveys. The advantages are obviously practicality and cost saving. The disadvantages are obviously possible biases due to assumption violations. Our philosophy of using a model based approach to estimating a non sampling error is not new and is now being widely applied to studying many other non sampling error problems.

3. A BRIEF REVIEW OF CAPTURE-RECAPTURE MODELS

It is obviously beyond the scope of this manuscript to review the extensive capture-recapture literature. For more information we recommend Seber (1982), White et al (1982), Pollock et al (1990) and Pollock (1991). Pollock (1991) is a review paper and a good lead into the literature and our treatment in this section follows it very closely. The other references are books and monographs for the serious reader with more time.

Here we briefly discuss the Lincoln-Petersen model for two samples, more general closed population and open population models for more than two samples, and finally a method which combines closed and open population models in one sampling design.

3.1 The Lincoln-Petersen Model

This is the oldest, simplest and best known capture-recapture model dating back to Laplace, who used it to estimate the population size of France. It was first used in fisheries by Petersen around the turn of the century. An excellent detailed discussion of this model is given by Seber (1982, Chapter 3).

In the original fisheries setting the method can be described as follows. A sample of M fish is caught, marked, and released. Later a second sample of n animals is captured, of which m are marked. An intuitive derivation of the estimator follows from equating the proportions marked in the sample and the population.

$$m/n = M/N \quad (3.1)$$

which gives

$$\hat{N} = Mn/m \quad (3.2)$$

A modified estimator with less bias in small samples is due to Chapman (1951) and is given by

$$\hat{N}_c = \left[(M + 1)(n + 1) / (m + 1)\right] - 1 . \quad (3.3)$$

An estimate of the variance of $\hat{N}_c$ is given by

$$\hat{Var}(\hat{N}_c) = \frac{(M + 1)(n + 1)(M - m)(n - m)}{(m + 1)^2(m + 2)} . \quad (3.4)$$

See for example Seber (1982, p. 60)

The crucial assumptions of this model are:

(a) The population is completely closed to additions and deletions,

(b) all the fish are equally likely to be captured in each sample, and

(c) marks are not lost or overlooked.

The assumption about closure can be weakened, but even for a completely open population where this estimator does not apply, a modification of the Lincoln-Petersen estimator is used. The assumption of equal catchability causes problems in most applications. There may just be inherent variability (heterogeneity) in capture probabilities of individual animals due to age, sex or other factors. There may also be a response to initial capture (trap response). In the next section, we consider closed population models with more than two samples that allow for time variation as well as heterogeneity and trap responses in the animals' capture probabilities. The loss or overlooking of marks can be serious. One way to estimate mark loss is to use two marks (Seber 1982, p. 94).

3.2 Closed Population Models

Closed population models require the assumption that no births, deaths, or migration in or out of the population occur between sampling periods. Therefore, these models are generally used for studies covering relatively short periods of time (e.g., trapping every day for 5 consecutive days). Capture histories for every animal caught

are the data needed for obtaining estimates under these models. Important early references are Schnabel (1938) and Darroch (1958), who considered models that assumed equal catchability of animals in each sample.

A set of models that allow capture probabilities to vary due to heterogeneity, (h), trap response (b), time variation (t), (i.e., capture probability for day i differs from that for day j) and all possible two- and three-way combinations of these factors is now available. The eight models [M(o), M(h), M(b), M(bh), M(t), M(th), M(tb), M(thb)] were first considered as a set by Pollock (1974) and were more fully developed by Otis et al. (1978), White et al. (1982), and Pollock and Otto (1983). Otis et al. (1978) provided a detailed computer program, CAPTURE, for use with their monograph. An updated version provides estimates for six of the eight models and a model selection procedure that aids the biologist in choosing a model. The model selection procedure is based on a variety of goodness-of-fits tests. Recently, Menkins and Anderson (1988) have emphasized that the model selection procedure is poor for small populations, unless the capture probabilities are unrealistically high.

3.3 Open Population Models

In many capture-recapture studies, it is not possible to assume the population is closed to additions and permanent deletions. The basic open population model suitable for this situation is the Jolly-Seber model (Jolly 1965; Seber 1965; Seber 1982; p. 196). The Jolly Seber model allows estimation of population size at each sampling time as well as estimation of survival rates and birth numbers between sampling times. Migration cannot be separated from the birth and death processes without additional information.

The Jolly-Seber model requires the following assumptions:

(a) Every animal present in the population at a particular sampling time has the same probability of capture,

(b) every marked animal present in the population immediately after a particular sampling time has the same probability of survival until the next sampling time,

(c) marks are not lost or overlooked,

(d) all emigration is permanent, and

(e) all samples are instantaneous, and each release is made immediately after the sample.

Assumptions (a), (c), and (e) were required under the basic Lincoln-Petersen model described in Section 2.2. Only marked animals are used to estimate survival rates to that, strictly, we do not need to assume equality of marked and unmarked survival rates. In practice however, the biologist will want to use the survival rate estimates to refer to the whole population. The Jolly-Seber model allows for some animals to be lost on capture and hence not returned to the population. The Jolly-Seber model also requires that all emigration is permanent. If animals emigrate and then return to the population they will have zero capture probabilities while absent. This so called temporary emigration is a serious assumption violation which can cause major bias in population estimates.

3.4 Combination of Closed and Open Models

Pollock (1982), Pollock et al (1990) and Kendall (1992) discuss sampling methods which allow the use of closed and open models in one design. One advantage of these methods is that it is possible to allow for unequal catchability whereas in the traditional Jolly-Seber model it is not possible to allow for unequal catchability. They also have the advantage of allowing for temporary emigration of animals.

3.5 Applications of Capture-Recapture Models

Capture-recapture models have obviously been widely applied to wildlife and fishery populations. A variety of novel nonbiological applications of capture-recapture methods have also now appeared. Several authors, including Wolter (1986, 1990) and Cowan and Malec (1986) applied capture-recapture to estimating the census undercount. Cowan, Breakey, and Fischer (1986) used it to estimate the number of homeless people in a city. Greene (1983) has used the method to estimate demographic parameters on criminal populations. Wittes (1974) and Wittes, Colton, and Sidel (1974) have used capture-recapture to estimate numbers of people with illnesses from hospital and other lists. The sampling of elusive human populations using cluster sampling, network sampling, and capture-recapture sampling was discussed by Sudman, Sirken and Cowan (1988).

4. USE OF CAPTURE-RECAPTURE MODELS IN THE LARGE PELAGIC SURVEY

The Large Pelagic survey is an angler survey conducted by the National Marine Fisheries Service and is basically a Telephone-Access Survey Design. A sample of fishing boat owners on a list are telephoned to obtain fishing effort information. Catch per unit effort information is obtained from a second sample of

boat owners at access points at completion of their fishing trips. The information from the two surveys is combined to estimate total effort and total catch of important species such as Bluefin Tuna.

A serious problem with this survey is that the list of boat owners used in the telephone survey is very incomplete. Therefore, classical sampling theory which assumes a complete frame of known size (N) is inadequate and has to be modified. The current method of estimating the size of the fishing boat list frame involves combining two lists, (a telephone list with a dockside list) and using the Lincoln-Petersen model. There are questions about whether this is the best approach. For example, it might be possible to combine more than two lists and if so then we could use the closed or open population models reviewed in Sections 3.2 and 3.3. However, we defer those questions and begin by reviewing and evaluating the current method.

4.1 The Lincoln-Petersen Model

4.1.1 Estimation of Frame Size (N)

Under the current method the "marked" boats (M) are those on the master list which is primarily derived from previous telephone interviews. The recapture sample is carried out dockside at gas pumps and the total number of boats intercepted or whose names are given by other captains (n) is checked to see which ones are "marked" (m) (i.e. on the original master list). Equations 3.2 or 3.3 can then be used to provide an estimator of the frame size (N).

$$\hat{N}_c = \left[(M+1)(n+1)/(m+1)\right] - 1.$$

Let us now consider the assumptions of this model and what effect violations might have on the bias of the estimator of N.

Closure

This assumption is likely to be violated. Fishing boats may be on the master list and then no longer take part in the fishery (losses). New fishing boats may join the fishery while it is in progress (gains). Ideally a separate estimate of frame size should be obtained for each two week time period. The advantage of using the Lincoln-Petersen closed model estimator is its simplicity and practicality. Biases in the estimator due to lack of closure could be either positive or negative.

Currently it is not known how the fishing fleet size is likely to change during the fishing season. A multiple capture-recapture sampling design would allow use of the Jolly-Seber model to estimate the fleet size during each period. Examination of these estimators and the survival rate and recruitment number estimators will enable us to evaluate the validity of the closure assumption. At the moment we can only make conjectures.

Equal Catchability

Violation of the assumption of equal catchability may be due to either inherent heterogeneity of capture probabilities between individuals or "trap response" where individuals that are marked have higher or lower capture probabilities than unmarked individuals. In either situation when the individuals on the lists are fishing boats we believe there is a potential for heterogeneity of capture probabilities among fishing boats. If heterogeneity is operating across both samples individuals "caught" on the first list will tend to be those with high capture probabilities and therefore they will more likely to be "caught" again on the second list. This means that the proportion marked in the second sample (list) will be too high and the estimator of N will be negatively biased. Note that this intuitive argument makes clear that is not heterogeneity per se which is the problem but that the heterogeneity continue across both samples. In other words, an individual's capture probability in one sample is very positively correlated with its capture probability in the other sample. Another way of stating the equal catchability assumption is that capture probabilities in the two samples are independent. One method of attempting to achieve independence of the capture probabilities in the two samples is to use totally different sampling schemes for the two samples. This is why we recommended earlier that one sample list be based on the telephone interviews and the other on dockside interviews. However, we do suspect that there is still another heterogeneity and lack of independence in capture probabilities. I believe that fishing boats which take a very active part in the fishery are more likely to be on any lists gathered (telephone or dockside). This heterogeneity will cause a negative bias on the estimate of frame size but we have no idea of the degree of this negative bias.

Marks Lost or Overlooked

The situation here is a little confusing. At first one might think that in this application there is not way that a mark could be lost or overlooked. However, this assumes that all boats have distinct names or that if boats do have the same name there is additional information like captains name which makes all individuals on the

lists unique. If there is any problem with lack of uniqueness it may not be clear whether a marked boat has been recaptured or not. Another related point is that agents may make errors in the records which make it hard to match up a recapture with the original record. A standard operating procedure is being developed and documented to minimize these kinds of errors in the future.

4.1.2 Estimation of Total Effort and Total Catch

Total Effort (E) is estimated by

$$\hat{E} = \hat{N}\,\bar{e} \qquad (4.1)$$

where $\hat{N}$ is the frame size (Fleet Size) estimate and $\bar{e}$ is the mean fishing effort obtained from the telephone sample. The evaluation of the properties of this estimator is more difficult than when N is known because both $\hat{N}$ and $\bar{e}$ are random variables. We suspect that $\bar{e}$ is biased high because fishing boats that do not fish much are less likely to be on the list. Unfortunately we cannot say that $\hat{N}$ will always be biased high or low. All three of the assumption violations discussed in 4.1.1 could be important (closure, heterogeneity, and mark loss) and it is not clear what direction the overall bias on $\hat{N}$ would take. The only possible approach is to use simulation with a variety of different scenarios for assumption violations. Using equation (2.4) we have the estimated variance of $\hat{E}$ is given by

$$\hat{Var}(\hat{E}) = (\hat{N})^2\,\hat{Var}(\bar{e}) + (\bar{e})^2\hat{Var}(\hat{N}) + \hat{Var}(\bar{e})\hat{Var}(\hat{N}) \qquad (4.2)$$

Total catch (C) is estimated by

$$\hat{C} = \hat{N}\,\bar{e}\,\bar{c} = \hat{E}\,\bar{c} \qquad (4.3)$$

where $\hat{E}$ is the estimated total fishing effort and $\bar{c}$ is the average catch per unit effort calculated from the dockside interviews. Properties of equation (4.3) are likely to be subject to similar concerns to equation (4.1). We do not know which direction bias in $\bar{c}$ is likely to go. Fishing boats that fish more often bias $\bar{e}$ high but do they also catch more fish biasing $\bar{c}$ high also? This may be a reasonable conjecture because the reason they fish more could be they tend to be more successful. Again, simulation may be helpful. We certainly cannot think of any other way of studying the properties of Equation (4.3). Using equation (2.4) we have the estimated variance of $\hat{C}$ i

$$\hat{Var}(\hat{C}) = \hat{Var}(\hat{E})(\bar{c})^2 + Var(\bar{c})(\hat{E})^2 + \hat{Var}(\hat{E})\,Var(\bar{c}) \qquad (4.4)$$

4.1.3 Illustration of the Method

In this section we present the frame size estimates, total effort and total catch for the Virginia Bluefin tuna fishery in part of 1992. These estimates are a part of a larger survey which covered the east coast of the U.S. from North Carolina to Massachusetts. The estimates are separate for charter boats and private boats.

Frame Size Estimates

Lists of unique private boats and charter boats were compiled mainly by telephone interviews from previous seasons. During the current 1992 season "marked" and "unmarked" boats were captured at gas pumps before or after fishing trips.

For private boats the list size was $M = 335$ boats before the season. A sample of $n = 374$ boats were contacted at gas pumps and of those $m = 49$ were marked. The Chapman estimator is

$$\hat{N}_c = \frac{(M+1)(n+1)}{(m+1)} - 1 = \frac{336 \times 375}{50} - 1 = 2519$$

$$\hat{Var}(\hat{N}_c) = \frac{(M+1)(n+1)(M-m)(n-m)}{(m+1)^2(m+2)} = \frac{336 \times 375 \times (335-49) \times (374-49)}{50^2 \times 51} = \frac{336 \times 375 \times 286 \times 325}{50^2 \times 51} = 91{,}856.4706$$

$$\hat{SE}(\hat{N}_c) = 303.08$$

Relative Standard Error $= 303.08/2519 = 0.12$

For charter boats the list size was $M = 47$ before the season. A sample of $n = 31$ boats were contacted at gas pumps and of those $m = 13$ were marked. The Chapman estimator is

$$\hat{N}_c = \frac{(M+1)(n+1)}{(m+1)} - 1 = \frac{48 \times 32}{14} - 1 = 109$$

$$\hat{Var}(\hat{N}_c) = \frac{(M+1)(n+1)(M-m)(n-m)}{(m+1)^2(m+2)}$$

$$= \frac{48 \times 32 \times (47-13) \times (31-13)}{14^2 \times 15}$$

$$= \frac{48 \times 32 \times 34 \times 18}{14^2 \times 15}$$

$$= 319.7388$$

$$\hat{SE}(\hat{N}_c) = 17.88$$

Relative Standard Error = 17.88/109

= 0.16

Total Effort and Catch Estimates

Total effort and total catch were estimated in weekly waves. Here we just illustrate the calculations for the week of the 8th to the 14th of June 1992.

Total Effort – Private Boats

$\hat{N}_c = 2519$ boats $\hat{Var}(\hat{N}_c) = 91{,}856.4706$

$\bar{e} = 0.15108$ trips per interview

$Var(\bar{e}) = 0.001242$

$SE(\bar{e}) = 0.0352$

Total Effort $= \hat{N}_c \times \bar{e}$

$= 2519 \times 0.15108$

$= 380.57$ trips

Variance (Total Effort) =

$Var(\bar{e})(\hat{N}_c^2) + \hat{Var}(\hat{N}_c)(\bar{e})^2 + Var(\hat{N}_c)Var(\bar{e})$

$= 0.001242 \times 2519^2 + 91{,}856.4706 \times 0.15108^2 + 91{,}856.4706 \times 0.001242$

$= 7880.9384 + 2096.6392 + 114.0857$

$= 10{,}091.6633$

SE(Total Effort) = 100.45

It is useful to also calculate the variance of total effort assuming that the frame size were known. In this case it is

Variance (Total Effort) = 7780.9384

SE(Total Effort) = 88.77

This shows that 89% of the standard error of total effort is due to variation in average effort and only 11% is due to estimation of frame size.

Total Effort-Charter Boats

Total Effort $= \hat{N}_c \times \bar{e}$

$= 109 \times 0.55$

$= 59.95$ trips

Variance (Total Effort) = 512.5100

SE(Total Effort) = 22.64

The variance of total effort assuming the frame size is known is

Variance (Total Effort) = 404.8926

SE(Total Effort) = 20.12

In this case again 89% of the standard error of the total effort is due to variation in average effort and only 11% is due to estimation of frame size

Total Catch – Private Boats

Total Catch $= \hat{N}_c \, \bar{e} \, \bar{c}$

$= 2519 \times 0.15108 \times 0.8276$

$= 314.96$ fish caught

Variance(Total Catch) = 21,820.1408

SE(Total Catch) = 147.72

Notice that catch estimates are very variable.

Total Catch – Charter Boats

Total Catch $= \hat{N}_c \, \bar{e} \, \bar{c}$

$= 109 \times 0.55 \times 6.7857$

$= 406.80$ fish caught

Variance(Total Catch) = 34,121.4219

SE(Total Catch) = 184.72

Notice that catch estimates are very variable and that the charter boat catch rate is much higher than for private boats.

4.2 More Than Two Lists

In Section 3 we indicated that there are a lot more modeling possibilities if one has multiple (greater than 2) lists. In this section we consider closed and open population models for the more general case.

We foresee the sampling scheme as follows. Before the start of the fishing season there would be a preliminary sample to establish a list (either telephone or dockside). During each time period (say two weeks) there would be an additional list compiled using a telephone or dockside survey. Now each individual boat would have a capture history which would indicate which lists it appeared on. (Suppose we have five

time periods then a capture history of 1 1 1 0 1 would indicate a boat appeared on the lists in all except the fourth time period).

The structure of the population would be as follows:

Marked Population Sizes

$M_0, M_1, \ldots, M_k$

Total Population Sizes

$N_0, N_1, \ldots, N_k$

The first question that has to be addressed is whether we need to use closed or open population models. The obvious way to proceed is to fit the Jolly-Seber open population model first and use it to evaluate the closure assumption.

4.2.1 Open Population Models

Under the Jolly-Seber model previously discussed in Section 3.3 the following parameters are estimable.

Marked Population Sizes

$M_0 \equiv 0, \hat{M}_1, \ldots, \hat{M}_{k-1}$

Total Population Sizes

$\hat{N}_1, \hat{N}_2, \ldots, \hat{N}_{k-1}$

Survival Rates

$\hat{\phi}_0, \hat{\phi}_1, \ldots \hat{\phi}_{k-1}$

Recruitment Numbers

$\hat{B}_1, \hat{B}_2, \ldots, \hat{B}_{k-2}$

Notice that it is possible to estimate the number of fishing boats in the fleet at each time in the season except the last (ie $\hat{N}_k$ cannot be estimated). One advantage of applying the model in this fashion with a preseason list is that any concerns with the preseason list due to it being out of date are taken care of by the model allowing for additions and deletions before the season begins. One disadvantage of the Jolly-Seber Model is increased complexity. Now each time period has its own frame size and there are also survival and recruitment parameters to estimate. Sometimes these parameter estimates have poor precision unless sample sizes are large. Another disadvantage of the Jolly-Seber model is that it does require the assumption of equal catchability.

Another important question about the use of the Jolly-Seber model is what is called "temporary emigration." A fishing boat might leave the fishery for some periods and then return. The Jolly-Seber model makes the assumption that fishing boats which leave do not return. This issue needs further investigation. Use of the robust design (i.e. combination closed and open models) allows for temporary emigration. This would necessitate having two lists obtained close together in each period.

4.2.2 Closed Population Models

If the Jolly-Seber model estimates of "survival" and "recruitment" suggest population closure the general closed population models reviewed in Section 3.2 could be applied here. The advantages are increased precision of $\hat{N}$ due to the use of more lists and increased robustness of $\hat{N}$ to unequal catchability. We believe these models need to be seriously considered. The disadvantage is primarily an increase in complexity.

5. DISCUSSION

5.1 Methods of Dealing with Incomplete List Frames

(i) Complete the List Frame

The advantage is that the survey researcher has a complete frame and does not have to generalize results for an estimated frame size. The disadvantage is the cost and possible impracticality of completing the list frame.

(ii) Use an Area Frame

The advantage is that one only has to enumerate the establishments in the areas to be sampled. The disadvantage is possible inefficiency if businesses are sparse in each large area.

(iii) Using List and Area Frame (Multi-Frame Approach)

The advantages are obviously increased precision and having all establishments covered. The disadvantage could be expense and impracticality.

(iv) Use of Capture-Recapture to Estimate List Frame Size

The advantage is having a practical method of lower expense than the first three approaches listed above. The disadvantages are potential bias if the assumptions of the capture-recapture method are violated and having to include variation due to frame size estimation in variance estimates of population total estimates.

5.2 Capture-Recapture Estimation of Frame Size

In this section we consider model assumptions, precision of estimates, estimation of population totals and the special problems in more complex sampling designs when the capture-

recapture approach to frame size estimation is used.

Model Assumptions

(i) Closure

Can the frame size be considered constant so that the closed population models be used? This will depend on whether the survey is just a snapshot at a single time point or whether a series of surveys over time are required. It will also depend on how quickly establishments go out of business and how quickly new ones arise. We suspect there will be the need for use of closed and open population models depending on the establishments being studied.

There is also the question of temporary emigration where establishments go out of the frame and then come back in again. This was considered a potential problem in the fishing boat example because boats could go inactive and then become active again. This may not be so much of a problem in other establishment surveys.

(ii) "Unequal Catchability" and Independence of Lists

As we discussed earlier ideally the lists used should be independent so that the estimates of frame size are unbiased. In practice it may not be easy to find two or more independent lists.

(iii) Mark Loss-Unique Identification of Establishment

Establishment names need to be unique and unmistakable or matches on different lists may be missed or mistaken. This was a problem in the fishing boat example in earlier years. We suspect this will not be such a big problem in most establishment surveys.

Precision of Estimates

The lists used need to be of sufficient size that the precision of the frame size estimate ($\hat{N}$) is adequate. Seber (1982, p. 96) discusses the Lincoln-Petersen estimate in detail and presents graphics of sample sizes required for various levels of precision. Pollock et al (1990) presents sample size information for the open population models.

Estimation of Population Totals

Once the estimate of frame size is obtained then that estimate will often be combined with a sample mean to obtain an estimate of a population total ($\hat{Y} = \hat{N}\bar{y}$). The estimate of population total is subject to possible bias and additional variance because $\hat{N}$ is estimated. The estimate may also be biased because $\bar{y}$ is not based on a random sample of the complete frame.

More Complex Sampling Designs

In this paper we have emphasized estimation of frame size in simple random sampling using the capture-recapture method. Further questions arise if more complex sampling designs are used. For example in stratified designs the question would arise of whether to estimate frame size in each stratum separately or to estimate the total frame size and then apportion it to the strata assuming equal probabilities of different strata on the incomplete lists. There is also the more complex question of how to estimate frame size in two stage sampling designs. This is obviously an area that needs future research.

REFERENCES

Chapman, D. G. (1951). Some Properties of the Hypergeometric Distribution With Application to Zoological Census, University of California Publication in Statistics, 1, 131-160

Cochran, W. G. (1978). Sampling Techniques (Third Edition), New York: John Wiley and Sons.

Cowan, C. D., Breakey, W. R. and Fischer, P. J. (1986). The Methodology of Counting the Homeless, In Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 170-175.

Cowan, C. D. and Malec, D. (1986). Capture-Recapture Models When Both Sources Have Clustered Observations, *Journal of the American Statistical Association*, 81, 347-353.

Darroch, J. N. (1958). The Multiple-Recapture Census I: Estimation of a Closed Population, *Biometrika*, 45, 343-359.

Goodman, L. A. (1960). On the Exact Variance of Products, *Journal of the American Statistical Association*, 55, 708-713.

Greene, M. A. (1983). Estimating the Size of the Criminal Population Using an Open Population Approach, In Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 8-13.

Hartley, H. O. (1962). Multiple Frame Surveys, Proc. Soc. Stat. Amer. Stat. Assoc., 203-206.

Hartley, H. O. (1974). Multiple Frame Methodology and Selected Applications. Sankhya, C36, 99-118.

Jolly, G. M. (1965). Explicit Estimates From Capture-Recapture Data With Both Death and Immigration-Stochastic Model.

Biometrika, 52, 225-247.

Kendall, W. L. (1992). Robust Design in Capture-Recapture Sampling: Modelling Approaches and Estimation Methods. Unpublished PhD. dissertation, North Carolina State University, Biomathematics Program.

Menkins, G. E., and Anderson, S. H. (1988). Estimation of Small Mammal Population Size. *Ecology*, 69, 1952-1959.

Otis, D. L., Burnham, K. P., White, G. C., Anderson, D. R. (1978). Statistical Inference for Capture Data on Closed Animal Populations. Wildlife Monographs, 62, 1-125.

Pollock, K. H. (1974). The Assumption of Equal Catchability of Animals in Tag-Recapture Experiments. Unpublished Ph.D. dissertation, Cornell University, Biometrics Unit.

Pollock, K. H. (1982). A Capture-Recapture Design Robust to Unequal Probability of Capture, *Journal of Wildlife Management*, 46, 752-757.

Pollock, K. H., Nichols, J. D., Hines, J. E. and, Brownie, C. (1990). Statistical Inference for Capture-Recapture Experiments, Wildlife Monographs, 107, 1-97.

Pollock, K. H, and Otto, M. C. (1983). Robust Estimation of Population Size in Closed Animal Populations From Capture-Recapture Experiments, *Biometrics*, 39, 1035-1049.

Pollock, K. H. (1991). Modelling, Capture, Recapture, and Removal Statistics for Estimation of Demographic Parameters for Fish and Wildlife Populations: Past, Present and Future. *Journal of the American Statistical Association*, 86, 225-238.

Schnabel, Z. E. (1938). The Estimation of the Total Fish Population of a Lake, *American Mathematical Monthly*, 45, 348-352.

Seber, G. A. F. (1965). A Note on the Multiple Recapture Census, *Biometrika*, 249-259.

Seber, G. A. F., (1982). The Estimation of Animal Abundance and Related Parameters (Second Edition), New York, MacMillan.

Sudman, S., Sirken, M. G., and Cowan, C. D. (1988). Sampling Rare and Elusive Populations, *Science*, 240, 991-995.

White, G. C., Anderson, D. R., Burnham, K. P., and Otis, D. L. (1982). Capture-Recapture and Removal Methods for Sampling Closed Populations, Los Alamos, NM: Los Alamos Laboratory.

Wittes, J. T. (1974). Applications of a Multinomial Capture-Recapture Model to Epidemiological Data, *Journal of the Statistical Association*, 69, 93-79.

Wittes, J. T., Colton, T., and Sidel, V. W. (1974). Capture-Recapture Method for Assessing the Completeness of Case Ascertainment When Using Multiple Information Sources, *Journal of Chronic Diseases*, 27, 25-36.

Wolter, K. M. (1986). Some Coverage Error Models for Census Data, *Journal of the American Statistical Association*, 81, 338-246.

Wolter, K. M. (1990). Capture-Recapture Estimation in the Presence of a Known Sex Ratio, *Biometrics*, 46, 157-162.