DECENTRALIZED COLLECTION OF MINING ESTABLISHMENT DATA WITHIN THE CANADIAN STATISTICAL SYSTEM

M. R. Dunn, T. Newman, J. Paquette, A.B. Siminowski
Energy, Mines and Resources Canada[1]
M. R. Dunn, Room 911, 460 O'Connor, Ottawa, Ontario

KEY WORDS: Decentralized, quality

Canada's national statistical system, recognized as a leading force in international statistical development, is largely centralized within the federal government agency, Statistics Canada. Under the Statistics Act, Statistics Canada is responsible to 'collect, compile, analyze, abstract and publish statistical information relating to the commercial, industrial, financial, social, economic and general activities and condition of the people'. It is also required 'to promote and to develop integrated social and economic statistics pertaining to the whole of Canada and to each of the provinces thereof and to coordinate plans for the integration of those statistics' [6].

To satisfy these responsibilities, Statistics Canada has developed and implemented a broad spectrum of establishment surveys covering the entire Canadian industrial base. These surveys have been designed to provide an accounting of the contribution of every sector to the economy. Nurtured by a statistical organization large enough to support major statistical initiatives, the creation of this survey structure has been paralleled in Statistics Canada by the development of 'leading edge' tools and systems to support its survey designs. These tools include the central Statistics Canada frame (the Business Register (BR)), computer software generalized to meet the varied needs of the agency's surveys, etc. The very size of the organization has led to its ability to create central pools of mathematical, statistical and technological expertise and to its capability to organize its internal operations to take advantage of the efficiencies engendered by large scale survey operations. Over the years, the Canadian statistical system has been regarded as perhaps the best in the world and this fact has been recently recognized in The Economist magazine (September 1991) [2].

In 1978, a decision was made to transfer the responsibility for a complex suite of monthly and annual establishment surveys for the non-fuel mineral industry (or the mining industry) from Statistics Canada to another federal government department, Energy, Mines and Resources (EMR). The transfer of survey responsibilities was made under the terms of an agreement between these two government departments. As part of this agreement, EMR also took responsibility for chairing and providing secretariat functions for the Federal-Provincial Committee on Mineral Statistics which coordinates the collection and dissemination of mining industry establishment information between the provincial mines ministries and the federal government departments interested in mining statistics (including both EMR and Statistics Canada). Responsibility for the official publication of the annual establishment data for the mining industry remained a responsibility of Statistics Canada. This unprecedented transfer continues to remain in effect at this time.

Since the time of the original agreement and transfer, arrangements under the agreement, including the provision by EMR of the range of statistics required by Statistics Canada for its System of National Accounts, have generally worked smoothly. Nevertheless, in 1992, largely to take into account shifting systems responsibilities, EMR initiated discussions aimed to review and update the agreement under which the original transfer was made. The decision to review this agreement raised the question as to whether it remained appropriate for these statistical survey activities to remain under the operational control of EMR, or whether it would be more appropriate for the surveys to be repatriated into the matrix of establishment surveys conducted by Statistics Canada. A key element in addressing this concern is the ability of EMR to provide statistical products and services of suitable quality. Is the quality of these products at the very least consistent with the general quality level of the other establishment surveys conducted by Statistics Canada?

[1] On June 25, 1993, Energy, Mines and Resources Canada became part of the federal department, Natural Resources Canada. The Mineral Policy Sector was subsequently renamed the Mining Sector.

It is thus necessary to put in place an effective strategy that will assure the quality of the establishment data produced. By continuing to fulfil its responsibilities under the agreement to the present, EMR already does take a number of positive measures to maintain the quality of its survey process. The data produced have been used by the System of National Accounts and, as mentioned earlier, cooperation between the two agencies has continued to the benefit of both. Nevertheless, it is necessary, within the current regime of increasing government restraint, to take advantage of the window of opportunity afforded by the review of the agreement between Statistics Canada and EMR to propose a series of measures, formed into a strategic plan that will assure the efficient delivery of quality information on the Canadian mining industry to its clients.

The goal of this plan is to identify measures that the organization external to the central statistical agency should take to ensure that the quality of its products are consistent with quality standards required by the national statistical system while, at the same time, to satisfy its own informational requirements. The challenge is to formulate a cohesive and effective strategy that will meet these twin goals.

By identifying the disadvantages of locating establishment surveys in a 'subject matter' organization outside the framework of the central statistical organization, specific actions or methods required to overcome or to minimize the perceived disadvantages can be isolated. At the same time the advantages of a decentralized statistical organization, once identified, can be enhanced to improve the value added to the statistical services and products provided.

It is the responsibility of the decentralized unit to ensure that a suitable level of quality for the data produced is achieved and maintained. This quality assurance must apply particularly rigorously to the data that are provided to Statistics Canada under the terms of the agreement and that becomes a non-trivial component of the System of National Accounts of that agency. But, on a more general plane, one must attempt to assure that the statistical information provided to the range of clients is of quality level fully acceptable to their needs. Information provided must meet the needs of our users for both historical and current data. These are not easy targets within the current limits of our financial restraints. Nevertheless they are targets which must be met.

The area within EMR responsible for the definition of the Canadian government's mineral and metal policy [3] is the Mineral Policy Sector (MPS) of the department. Part of that policy calls for the provision of timely and accurate information on the mineral and metal industry in Canada. To assure this, MPS is drawing together a program that will assure the quality of its statistical products and services.

The basic activities in this program include:

(1) Maintaining and improving its statistical processes. A cornerstone to any quality assurance program must be to seek to improve the statistical quality of its data in terms of minimizing traditional survey errors (both non-sampling and, where appropriate, sampling errors). These 'traditional' survey errors are discussed in standard literature such as Fellegi and Sunter[4]. Methods of compensating for or evaluating each type of error are provided in the Statistics Canada publication, Quality Guidelines [7] which can be used as a guide in this process. Although MPS has had a program designed to consistently and constantly edit survey responses and validate the data produced on an ongoing basis, there is a need to restructure this program into a cohesive and efficient editing scheme involving both electronic procedures and specialist review.

(2) Using the policies, standards and methods developed by Statistics Canada as a guide in improving the quality of our process. By the selective imitation of the policies, standards and methods that the central agency has set in place to ensure the high quality of its products and reputation, EMR can share the benefits accrued by using these policies, etc., as models where this is feasible.

(3) In addition, undertaking a process of 'benchmarking' its operations against the central statistical organization in a way similar to the 'benchmarking' exercises described in Almdal [1]. In this way, operational efficiencies can be identified and harnessed. Likewise by participating in international data review processes and by ongoing communication with the international fraternity of organizations involved in parallel survey activities, our data assurance measures can be consistent with international standards.

(4) Utilizing highly-qualified personnel and systems available off-site. EMR proposes to take advantage of the resource pool available at, for example, Statistics Canada by means of personnel exchanges and agreements to jointly sponsor projects of mutual interest. It is hoped that these efforts will contribute to the development of more extensive bank of internal statistical expertise. In addition we wish to take full advantage of the training opportunities available through

the central statistical agency.

(5) Forging fair and workable agreements with our partners in Statistics Canada and the ministries in the provincial governments, leading to the minimization of duplication of activity between government agencies.

(6) Reinforcing communication links with our statistical partners, Statistics Canada and the provincial ministries. A primary vehicle for promoting these communication links is the Federal-Provincial Committee on Mineral Statistics, part of the network of consultative committees established by Statistics Canada.

(7) Vigorously pursuing feedback with our total client base to ensure that their needs are being properly addressed and the products and services provided are of sufficient quality. This feedback will be gathered by both systematic and periodic surveys of, and informal interactions with, our various client bases.

(8) Establishing an overall strategy for the dissemination and evolution of survey products. Part of the Sector's communication strategy involves continuing opportunities for client feedback. The Sector has initiated planning aimed at recovering costs for certain products and services that it provides.

(9) Making senior management aware of the real benefits, needs and costs of a statistical program on an ongoing, and not only crisis, basis. The visibility of the statistical program should be promoted both inside and outside the host organization.

(10) Continuously looking forward
..to a full structural review of our surveys with the central agency
..to participation in the review and amendment of common classification systems, and
..to respond to the evolving nature of both the statistical system and the governmental structure in which it finds itself with a program of continual improvement.

The Mining Industry Frame

In most statistical surveys, it is absolutely critical that the frame chosen is a close approximation of the target population. Moreover, in the context of the mining industry establishment surveys, it is important that the frame is consistent in its treatment of companies across all establishment surveys and industrial groupings. For example, mining companies will often operate both production establishments (producing mines) and refining or smelting establishments. The first set of establishments is surveyed by EMR, the second is the survey responsibility of Statistics Canada. Under these circumstances it is necessary that the parties cooperate to ensure not only that the establishment coverage is complete and that updates are exchanged but also that the collection and the editing of the survey data are harmonized, thus removing duplication of collection cost and minimizing the data collection burden on the respondent.

For the most part, the mining industry consists of large and long-term companies. These are readily identifiable and coverage of this population is considered to be virtually complete. This fact is verifiable by the commodity specialists who will quickly notice the impact of the omission of even one company in a total or a count. However, certain sub-populations, covered by the mining industry surveys, particularly the population of sand and gravel pit operators, are highly volatile not just in terms of the length of their operating existence but also in terms of the industrial classification to which they may be assigned. This year's farming operation may become next year's gravel pit operator and then may revert to farming again in the succeeding year and so on..

Statistics Canada operates the Business Register, a comprehensive business survey frame, covering all establishments and other statistical units of interest across the complete range of industrial classifications. The Business Register therefore has been developed by Statistics Canada as a frame for certain of its own surveys of the mining industry. Its monthly employment survey, the Survey of Employment, Payroll and Hours (SEPH), covers a cross-section of Canadian employers including those in the mining industry. In order to ensure that estimates from the respective surveys can be consistently used by industry analysts, it is necessary to assure that the frames and the associated classification structures used as a basis to survey the populations are similar and compatible.

EMR and Statistics Canada will develop a protocol for the exchange of updates. In its agreements with each the provinces with mining industries, EMR will continue to collect and exchange frame updates with each participating province. By systematizing this information flow, the various partners in the collection of data on the Canadian mining industry can be assured that the frame information is both reasonably up to date and consistent with that used by all parties to the interchange agreements. Statistics Canada has also agreed to initiate the process of reconciling the current EMR frame with the BR within the next year.

Quality of the Data Produced

It is worth noting at the outset that the transfer of the survey responsibility to the decentralized organization brought with it no known direct deterioration of data quality. Response rates to survey questionnaires remain at well above 90%. Moreover, it is our intention to collect and retain a greater database of quality measurement information. Contacts established by MPS at various levels of the corporate structures of our major respondents have, we believe, improved the quality of data provided and will aid in the development of restructuring of the data collected as the industry itself evolves. This relates particularly to the growing importance of the use of secondary materials in the mining industry and the need to capture information related to the environmental initiatives. Another area in which work is currently being undertaken is the categorization of exploration expenditures under the aegis of the Federal-Provincial Committee.

Data on mines openings and closings, and the impact on industrial employment, collected in MPS are being compared with changes in employment levels generated by the SEPH at Statistics Canada to ensure that these data sources show consistent trends. Any discrepancies noted can be compared. Similarly data collected through the industrial surveys will be compared for consistency with Statistics Canada employment levels.

With the advent of new statistical data systems, increased efforts are being made to develop electronic databases which can be shared among the data-gathering partners. This will collectively reduce data capture costs, minimize the joint data editing costs and provide an opportunity to fully ensure the production of consistent statistical tabulations between the parties involved.

In order to control workloads on personnel including those workloads shared with the provinces in the face of fiscal restraint, we expect to introduce sampling methods and other survey design measures into what had formerly been full census coverage. This will naturally add a sampling error component to our survey errors, a component which will be fully estimated and documented. A greater effort will be made to control respondent and data capture errors through a systematic approach to electronic editing, through a revitalization of the current editing notes and perhaps through the introduction of selective editing processes within the framework of this system. Interest is being taken in the selective editing processes being developed at Statistics Canada and discussed at this meeting.

Primary Statistical Analysis

In order to make the issue of data quality visible, a data quality statement will be phased into each regular publication used as a primary release mechanism for the data produced. Statistics Canada maintains a clear and concise policy on data quality statements that appear in its releases. The first stage of this initiative will be the preparation of a quality statement consistent with Statistics Canada policy in all Statistics Canada releases in which EMR provides data: the second stage will be the satisfaction of this requirement for all systematic data releases of the EMR statistical unit.

Analyses of the impact of the Canadian mining industry on the Canadian economy often involves not only the analysis of the mining industry itself but also the industries directly dependent on that industry or downstream. These industries include the smelting and refining and metal semi-fabrication and fabrication industries. Since data for the downstream industries are collected by Statistics Canada, it is necessary to build analyses from the data sources from the two agencies. It is also necessary, and this is a very current concern, that we develop the industrial structure profile in concert with Statistics Canada and other statistical organizations undertaking similar analyses.

Finally an ongoing program to seek feedback from our client base and indeed to periodically redefine that client base is absolutely essential in defining the statistical products and services that we produce. This comes not only through a systematic and continuing program of client surveys and consultation but also through day-to-day contacts with Statistics Canada, our own commodity officers and policy analysts and meetings with our provincial and international counterparts.

In a paper prepared for the ISI in 1975, Norwood [4] concludes that, provided objectivity and impartiality are maintained, the organization which produces the statistics should indeed analyze them. While Statistics Canada holds its objectivity and impartiality as a keystone value to its operations, it may be difficult to assure the general public that other line government departments can maintain the same overall level of objectivity in their operations, as they are often seen as advocates of an industry or an issue. It is a continuing challenge for MPS to be seen as objective and impartial in its construction of its statistical outputs, including any data analyses that it conducts.

Conclusions and Directions

It is the conclusion of this review that the placement of a survey unit responsible for the collection and dissemination of industrial establishment statistics outside of the central agency presents a viable alternative to its location within the administrative umbrella of that agency. In fact, there can be major, perhaps over-riding, advantages to the organization hosting that unit, for the responding business and for the national statistical system as a whole. But it is incumbent upon the decentralized survey organization to operate under a strategy that will assure that the quality of its statistical contributions are subject to standards essentially as rigorous as those observed by the central agency. As a consequence, it is also important that the decentralized unit take advantage of the expertise, experience and facilities of the central agency to meet that goal.

Given good will and sound communication links between the organizations involved, and resourcing to allow suitable quality assurance activities to be supported, the maintenance of decentralized establishment data collection activity is indeed possible and, in some instances, a preferable option. Indeed the central agency can benefit by existence of small decentralized units to provide a positive, and perhaps critical, input into its activities.

However, it remains crucial that resources in the organization hosting the decentralized establishment statistics activity be sufficient to support an adequate quality assurance program. And it is important to recognize that, without the leadership and investment of the central agency, the success of the decentralized unit would be compromised.

REFERENCES

[1] Almdal, Bill, Continuous Improvement with Benchmarking, Paper presented to the 95th Annual General Meeting of The Canadian Institute of Mining and Metallurgy (CIM) - May, 1993

[2] Economist, The, September 1991

[3] Energy, Mines and Resources Canada, The Mineral and Metal Policy of the Government of Canada, May 1987

[4] Fellegi, I.P. and Sunter, A.B., Balances Between Different Sources of Survey Errors - Some Canadian Experiences, Sankhya, 36 Series C, (1974), pp. 119-142

[5] Norwood, Janet L., Should Those Who Produce Statistics Analyze Them? How Far Should the Analysis Go?, Invited paper, ISI, Warsaw, 1975

[6] The Statistics Act, An Act respecting statistics in Canada,
Office consolidation, March 1992

[7] Statistics Canada, Quality Guidelines, Second Edition, 1987

IMPROVEMENTS TO FRAME DATA QUALITY IN THE SMALL BUSINESS SECTOR

M. Charron-Corbeil and N. Falardeau, Systems Development Division
R.H. Coats Bldg, 12th floor, Tunney's Pasture, Ottawa, Ontario K1A OT6, Canada

Key Words: Small business frame, establishment-based surveys, automatic updates.

1. Introduction

The Statistics Canada Business Register is a central repository of information on approximately 1.5 million business entities operating in Canada. It is used as a list frame from which annual and sub-annual establishment-based surveys select their universes and samples. Small businesses account for only about 15% of Canadian economic activity but they represent about 92% of the population on the Register. Keeping this very large number of businesses up to date depends heavily on automatic processing. The algorithms used by these automatic processes must be continually monitored and refined to ensure the highest possible quality in the data supplied by the frame.

In 1992, a number of enhancements were implemented by Statistics Canada to the systems maintaining the small business sector, with the objective of improving the quality of establishment data. Two stratification variables of interest to surveys dealing with these businesses were targeted: Gross Business Income (GBI) and Employment Size (ES).

The purpose of this paper is to describe recent modifications made to the automatic update procedures for these two variables in terms of both concepts and practical issues of implementation, and how they have improved the quality of the data on the frame. Results illustrating the improvements are included.

2. Gross Business Income

2.1 Concepts

The GBI variable serves a dual purpose on the Business Register list frame. It is used as a stratification variable for annual and sub-annual surveys and it is used to identify potentially large businesses in the real world, i.e. businesses whose GBI is above a certain pre-determined upper threshold for its geography/industry classification for an extended period of time. Such units are further investigated and profiled to ascertain their class membership (large or small).

A GBI value is obtained through a model that was conceived in 1986 as part of the Business Survey Redesign Project. Several alternatives were considered; the current GBI model was adopted mainly because it was operationally appealing and relatively easy to implement. Its functional form is :

$$GBI_i = R * NI_i * \frac{SUMREM_i}{M_i} * 12 ,$$

where R is a ratio of total annual operating income to total wages and salaries, NI is a ratio of total wages and salaries to total remittances and $SUMREM$ is the sum of remittances in the past M months, where M is at most twelve. The subscript i denotes the i^{th} record.

The GBI model uses administrative data to produce an estimate of the annual operating income for small business remitters with employees. Remittances are obtained for approximately 900,000 unique Payroll Deduction (PD) accounts on a monthly Revenue Canada Taxation (RCT) file. The model also uses R, NI *or* I (default used when NI is unavailable) ratios computed using tax data that are available on a delayed basis, i.e. source files may be anywhere from 16 to 24 months out of date. We do not consider this to be a major drawback because ratios have a tendency to be stable over time. Extensive testing has confirmed this assumption. As well, the ratios are computed in a robust fashion; median classification cell ratios are used for R and I. The NI ratios are computed at the PD record level and are susceptible to anomalies in remittance patterns of individual businesses. This is a desirable characteristic that allows GBI to retain its individuality.

2.2 Implementation before 1992

The GBI model is executed on a monthly basis within a process called PAYDAC. Before a GBI value can be estimated, a small business must satisfy two conditions: its sum of remittances for the last 12 months must be greater than $0 and its average monthly remittances (AMR) must be greater than $227.34 - the lowest possible amount that a business can remit and be considered a large business for any industry/province combination. Ideally, all businesses that meet these conditions should automatically be updated and the GBI on the frame should be current in order to make stratification and estimation as efficient as possible.

Due to time and cost constraints associated with the production system, only a fraction of all records can be updated each month. To determine which records are to be updated, a number of GBI ranges were created to group units into different size strata. A business is refreshed only if its new GBI displaces it into one of the neighbouring strata. This practice gives rise to either downward or upward bias in the GBI. For example, in an inflationary situation, the GBI has a tendency to increase. It may do so for several months without moving up into the next size stratum and being refreshed. Thus the GBI remains artificially low. It may eventually cross the stratum boundary and be updated at which time the resulting jump reflects an accumulation of monthly increments and obscures the true movement in the GBI. The reverse would be true during recessionary times.

2.3 Improvements to the existing model

To achieve the objective of correcting the unacceptable level of overestimation or underestimation, a study was done in late 1990 from which a report entitled "GBI Model Evaluation" (Patak and Whitridge) was produced containing a list of recommendations of which three were adopted:

- GBI be computed for all records regardless of the AMR.
- The ratio tables be updated annually.
- Change the way we determine if an update should take place by removing range checking and introducing a technique to apply the most significant updates using a percent change approach.

2.4 Implementation since 1992

An improvement of the GBI value on the Business Register was immediately visible following the removal of the AMR verification from the algorithm. It was feasible to implement this recommendation without any delay after its approval because it required minimal system changes.

The ratio tables are updated annually as the tax-based files are made available. The R and I ratios represent location estimates (means or medians) at the geographical (province) and economic activity (major industry division or three digit Standard Industrial Classification) level. In the previous version of the ratio tables the location estimates were often based on fewer then 20 observations. This led to an increased variability in the ratios from year to year and the occurrence of outliers.

In the current version of the tables all R and I ratios are computed using at least 25 observations to reduce local and temporal variability. The principles of exploratory data analysis and smoothing have been employed to safeguard against large shifts in the location and scale parameters of the underlying distributions caused by anomalies in the economy. To stabilize GBI estimates and help avoid unnecessary jumps in the series of implementation, outlier detection was incorporated in the software.

To implement the last recommendation, an index originally proposed by David Birch (1987) was chosen for three reasons: (i) it avoids updating records that change by a small amount if they cross a GBI range boundary, (ii) it ensures that small firms which have a small absolute change in GBI are not solely updated on the basis of percentage change, and (iii) it allows larger firms to be updated due to a substantial absolute change even if the percentage change in their GBI is below the cutoff value. The formula for the index is the product of absolute and percentage changes in GBI and is as follows :

$$\textit{Birch Index} = \frac{(\textit{GBI model} - \textit{BR GBI})^2}{\textit{BR GBI}}$$

Supplementing the GBI model with the Birch Index was incorporated in the production cycle starting February 1992. The monthly process is executed in several steps: (i) creation of a Birch Index cut-off table, (ii) selection of a cut-off value for a given month, (iii) execution of the GBI model, and (iv) updating of the Business Register.

The Birch Index cut-off table identifies for each percent value from 1 to 99, a cut-off point and the number of expected updates. The table is created in three steps. First, a Birch Index value is calculated and stored on a file for each eligible PD record. Exceptions such as (i) large businesses, (ii) businesses with tendency changes, (iii) businesses whose GBI value on the BR is equal to zero, (iv) businesses whose GBI value on the BR is non-zero and is to be set to zero, (v) and new businesses (birthed that month), are ignored. Second, the number of observations is tabulated. Third, the number of updates is calculated by multiplying the percent (1 to

99) by the total number of observations minus the exceptions. We obtain the cut-off value for the specific percent using the Birch Index of the PD record whose rank matches the number of updates to be performed. To predict an accurate number of updates, this table must be recalculated monthly.

The choice of the cut-off value for a given month is a balance between the tightness of the constraint for the purpose of quality and the number of updates considered affordable. Although it is possible to select a Birch Index value that detects a certain degree of change in GBI, time constraints of the monthly PAYDAC production must be considered. In order to control the impact of the Birch Index in terms of elapsed computer time, it is important that the index not only target which records will be refreshed but also operate as a management tool to control the volume of updates.

The GBI model is executed for each small business if the sum of its remittances for the last 12 months is greater than $0.

For each qualifying record, the Birch Index is calculated and the result is compared to the chosen cut-off value. The GBI for a business will be refreshed if the Birch Index is equal to or greater than the cut-off value.

2.5 Results

2.5.1 Identifying large businesses

The improved GBI has shown an increased level of resolution resulting in almost a 40% reduction in the number of business entities identified for profiling.

Large businesses identified with old GBI	74,456
Large businesses identified with new GBI	34,765

2.5.2 Quantitative comparison of updates

The following table summarizes the number of updates applied using GBI ranges versus Birch Index.

If we compare the "%" columns, we see that there is no clear pattern for Ranges, whereas BI shows a greater percentage of the base as GBI increases. In addition, we know that those updates invoked by BI are due to significant changes in terms of both absolute and percentage change in GBI.

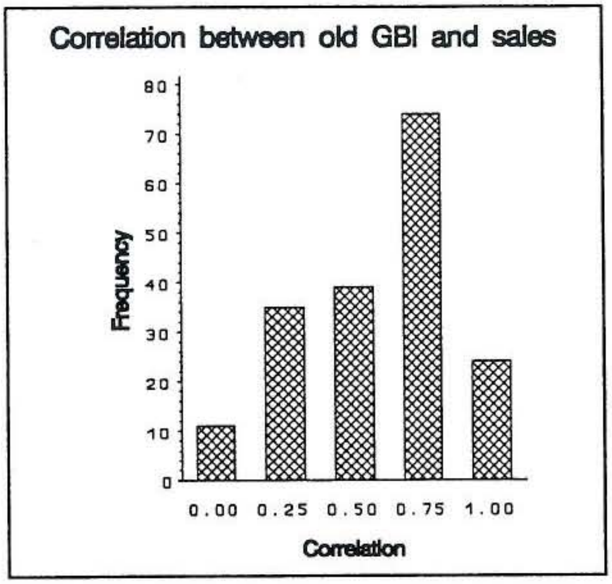
GBI Range ($)	Ranges (#)	Ranges (%)	BI (#)	BI (%)	Total Compared
1, < 10K	9,859	13	9,573	12	77,525
10K, < 25K	220	0	13,437	16	85,707
25K, < 50K	2,569	2	18,934	16	118,535
50K, < 100K	8,386	6	28,071	20	143,256
100K, < 200K	12,592	8	36,239	24	152,262
100K, < 250K	10,853	24	12,008	26	45,347
250K, < 500K	11,616	10	33,024	27	120,698
500K, < 750K	8,198	16	15,229	29	52,428
750K, < 1M	6,186	22	8,645	31	28,012
1M, < 2M	4,230	11	14,010	35	39,757
2M, < 5M	2,010	10	10,554	55	19,362
5M, < 10M	801	22	2,920	79	3,697
10M, < 15M	337	52	559	86	652
15M, < 20M	147	64	228	100	228
20M, < 25M	95	98	94	97	97
25M, < 50M	59	39	136	89	152
50M, < 75M	26	74	35	100	35
75M, < 100M	8	62	13	100	13
100M, < 500M	6	33	18	100	18
500M+	0	0	1	100	1

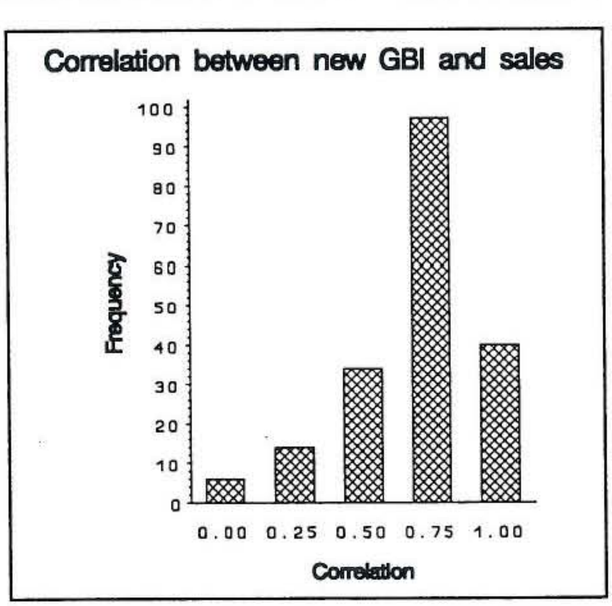
2.5.3 Stratification

A business - or an establishment - survey collects information that is germane to one or more sectors of the economy. This may be sales for MWRTS (Monthly Wholesale and Retail Trade Survey), the value of shipments for MSM (Monthly Survey of Manufacturing) or the number of employees for SEPH (Survey of Employees, Payroll and Hours). Often there exists a proxy variable available on a survey universe basis that is correlated with the variable of interest.

For pure retail operations GBI is assumed to be a good predictor of sales and as such could be used as an auxiliary variable to design a three-variable (geography, trade group and size) stratification scheme. If the correlation between GBI and sales is fairly high, stratification on GBI can be almost as good as with using the variable of interest. Preliminary testing shows that the new GBI is indeed a better predictor of sales than the old GBI and substantial gains in assigning units to their correct size strata can be realized.

The following plots provide a summary of the correlation between sales and the two sets of GBI.





The new GBI is visibly a better proxy for sales. The strengthening of the relationship between the two variables allows not only for a much improved estimation but also for a more efficient stratification of the MWRTS sample. This can be attributed to several improvements in the creation of the *R*, *I* and *NI* tables used as inputs to the GBI equation. Observations deemed outlying have been removed from the *NI* table and new methodology has been put in place to ensure its stability over time.

3. Employment Size

3.1 Concepts

The employment Size stored on the Statistics Canada Business Register represents the maximum employment observed over a 12-month period. The major user of the Business Register Employment Size information is the monthly Survey of Employment Payroll and Hours (SEPH). The survey requires the monthly Business Register updates as input into selecting and maintaining its sample.

3.2 Implementation before 1992

Before December 1992, Employment Size was initialized at the time of new employer registration or by a default value of "1", if the registration employment information was not available and the business showed a GBI greather than $0. Thereafter, the employment size was updated only for businesses contacted through one of the six surveys reporting employment data. Only 13% of the small business sector is contacted each year. By not having up-to-date employment data on the frame from which the sample is derived, the statistical efficiency of the survey allocation and sampling processes is adversely impacted. The under or over allocation of units to the strata increases as the frame data ages.

3.3 Implementation since 1992

In December 1992, the Employment Size Coding function was implemented on the Business Register. The objective was to improve and maintain the overall validity and reliability of Employment Size data for small businesses. This was achieved by adding to an existing automated monthly process, the estimation of a number of employees for each small business in conjunction with an updating mechanism which protects the data obtained through direct respondent contact as well as restricts the number of updates to the Register.

3.3.1 Estimation of number of employees

All employers in Canada must have at least one Payroll Deduction account in order to remit monies for Income tax, Canada Pension Plans Contributions and Unemployment Insurance premiums. Revenue Canada and Taxation provides that information each month to Statistics Canada to maintain the Business Register.

In December 1992, the existing monthly Payroll Deduction account process (PAYDAC) was changed to estimate the number of employees for small businesses. The formula used to estimate the Employment Size is :

$$\hat{ES}_i = \frac{NI_i * CMR}{AWE * 4.35},$$

where ES_i is the estimated employment size for the ith Payroll Deduction (PD) account, NI is a ratio of total wages and salaries to total remittances corresponding to the account, (if that information is unavailable a corresponding ratio at the Standard Industrial Classification (SIC) and province level is used), *CMR* is the current month remittance, *AWE* is the average weekly earnings for the month, (SEPH supplies that information at the SIC and province level), and 4.35 is the average number of weeks in a month. Each month the estimated value is stored on a historical file where 24 months of estimated values are kept for each active payroll deduction account.

The reliability of the model (columns) has been verified against SEPH reported data (rows). The following figure shows, for the same 12-month period, that 77% of the small businesses compared were classified in the same stratification range. A higher percentage of discrepancy in some cells (*) can be explained by the fact that some smaller firms have a tendency not to remit on a monthly base. More than one month of remittances registered in one month will generate, for that particular month, a higher employment size.

SEPH ranges	0	1-19	20-49	50-199	200+	SEPH total
0	307 0.91%	235 0.69%	14 0.04%	6 0.02%	0 0.00%	562 1.66%
1-19	3,068 *9.06%	22,117 65.29%	931 2.75%	123 0.36%	5 0.01%	26,244 77.47%
20-49	210 0.36%	1,449 *4.28%	2,366 6.98%	393 1.16%	6 0.02%	4,424 13.06%
50-199	123 0.36%	137 0.41%	837 *2.47%	1,375 4.06%	23 0.06%	2,495 7.36%
200+	31 0.09%	2 0.01%	9 0.09%	93 0.27%	18 0.05%	153 0.45%
Model total	3,739 11.04%	23,940 70.67%	4,157 12.27%	1,990 5.87%	52 0.15%	33,878 100%

3.3.2 Updating mechanism

Not all small businesses are updated each month. A set of conditions must be met before an update can be performed. A list of all small businesses contacted, either by phone or through a survey during the last 12 months, is accessible. If the company has been contacted and the Employment Size reported is not zero then the company is not updated, since the employment size obtained by contact is recorded on the Business Register. If the reported value is zero or if no contact occurred in the last 12 months, the process checks the registration information. This means that, the registration information must not be older than 12 months and it must have been captured within the first year of the appearance of the business on the Business Register. If the registration information meet these conditions, and the registration value for employment is not zero, the company is not updated.

If the process determines that the company can be updated, using the historical record, related to the company, the process selects the second maximum Employment Size estimated in the last 12 months. The methodologists and SEPH representatives agreed on using the second maximum value instead of the maximum, to avoid the selection of a figure resulting from an abnormality in the remittances for a given month (such as a strike). The second maximum employment size is then compared to the one already on the Register. The two figures must be in different pre-defined ranges for the company to be updated. Detection of significant changes using ranges became necessary in order to reduce the number of register updates.

Any business having its employment size going from less than 200 to more than 200 will automatically be signaled for manual inspection. 200 employees and over are self-representing units in the SEPH survey, where the ones with less than 200 are weighted units.

3.4 Results

Before running the new Employment Size Coding function for the first time, a special process identified all the Payroll Deduction accounts having no remittances for the last 12 months and updated their Employment Size to "0". This special process was necessary to avoid a huge amount of updates in the regular monthly process. The first monthly process, including the new function, cleaned-up all others. 346,361 accounts were updated. The following table shows their distribution. The two first ranges were mostly affected by the override of the default value of "1".

Ranges (for updates)	Jan -Dec 1992 (old process)	Jan - Dec 1992 (new process)
0	11,459	95,152
1 - 4	179,497	75,551
5 - 9	86,157	85,322
10 - 19	45,734	54,262
20 - 49	18,702	26,843
50 - 99	3,997	6,188
100 - 199	743	2,130
200 - 499	43	741
500 - 999	13	119
1000 - 1499	4	16
1500 - 2499	2	20
2500 - 4999	5	12
5000+	5	5
	346,361	346,361

4. Conclusion

The results validate the premise that the GBI index reduces the tendency for small firms to be updated solely due to a large percentage change in their GBI when the absolute change in GBI is insignificant. The Birch Index favours larger firms whose absolute change in GBI is substantial.

The correlation between the new GBI and other variables of interest to business surveys, has improved enough to for GBI to be used beyond stratification of samples. Other uses of GBI such as producing raking estimates are currently under investigation.

By the implementation of the new Employment Size function, we improved the data quality of our small business universe. Many businesses had not been updated since the loading of the register in 1987. When the new function was installed the employment values of approximately 500,000 small businesses were set to 0, because they had shown no remittances within the previous 12 months. 346,000 had a range change and 180,000 were not updated because they has been contacted within the last 12 months. The remainder showed no stratum change. Since the installation, the number of employment range changes averages 35,000 to 40,000 per month. More up-to-date Employment Size data allows the users to get more optimal samples and hence more efficient survey estimates.

5. References

Birch, D. (1987). Job Creation in America. The Free Press. New York.

Charron-Corbeil, M. and N. Falardeau (1992). Business Register GBI Coding Final Requirements. Statistics Canada Technical Report.

Hutchinson, D., Z. Patak and B. Chun (1992). Strategy to Implement a new GBI Update Criteria within the New BR - Testing and Analysis. Statistics Canada Technical Report.

Patak Z. and P. Whitridge (1991). GBI Model Evaluation. Statistics Canada Technical Report.

Patak Z. and D. Hutchinson (1992). Enhancements to the GBI Model Ratio Tables. Statistics Canada Technical Report.

Charron-Corbeil, M. and G. Hamel (1992). Business Register Employment Size Coding Final Requirements. Statistics Canada Technical Report.

Lundin B. (1992). Methodology and Systems Documentation for the Employment Size (ES) Process and the Gross Business Income (GBI) Process. Statistics Canada Technical Report.

AUTOMATIC GENERATION OF STANDARDIZED STATISTICAL STRUCTURES IN THE STATISTICS CANADA BUSINESS REGISTER

M. T. Barfoot, Statistics Canada
"12-R" R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6

KEY WORDS: Business Registers, Computerization, Economic Surveys, Standardization, Statistical Units, Temporal Database.

1. INTRODUCTION

1.1 The Statistics Canada Business Register

The Statistics Canada Business Register is a database containing data relating to the universe of businesses in Canada. In the context of this database the term business includes governments, non-profit organizations, etc. The Business Register is used as the frame for the major establishment surveys in Canada. Businesses stored on the database are placed into one of three distinct groups; the Integrated Portion (IP), the Non-Integrated Portion (NIP) and Insignificant businesses (ZIP). This placement is based on the Gross Business Income (GBI) of the business as well as the type of activity it is engaged in and the province within Canada in which it is physically located. The IP group contains the 10% of Canadian businesses (about 130,000 establishments) which account for around 80% of the gross national product. The businesses in the NIP group account for approximately 20% of the gross national product and consist of 50% of the businesses on the Business Register (about 770,000 establishments). The ZIP group contains the remaining 40% of businesses that have an insignificant contribution to the gross national product.

The Business Register is also a temporal database, which means that all states that a business has ever had are stored in the database. Each database update is associated with a double time stamp that represents the real-world effective date and the date on which the update was made to the database. The Statistical structure generator must take this temporal aspect into consideration. This paper deals with the generation of Statistical structures in the IP of the Business Register.

1.1.1 Why are Statistical Structures Needed?

At Statistics Canada economic data is collected from the real-world using a variety of methods. Statistical structures on the Business Register are used to transform the variety and complexity of real world business structures into a standardized four-level model that matches the requirements of the economic survey program (such as sampling and targets for data collection).

1.1.2 Legal, Operating, Statistical, Administrative and Collection Structures

The data in the IP of the Business Register are stored in five different types of structure: Legal, Operating, Statistical, Administrative, and Collection. The individual units within a structure are called "entities." Legal structures represent the legal ownership and/or control of a business by corporations and individuals. Operating structures represent the organizational and accounting structure of a given business. Statistical structures are a standardized, four-level representation of Operating structures that allow surveys to view the broad range of business structures in a uniform way from the Statistical perspective. Administrative data consist of the tax records associated with a business. Collection entity structures represent the reporting arrangements that various surveys have with their respondents.

Statistical structures are generated from Operating structures using a computer based system. The Operating structure is a representation of how a real-world business sees itself in terms of organizational entities managing production entities. Production entities are at the lowest level in the Operating structure and represent single geographic locations (i.e., plants, warehouses, retail outlets) where goods or services are produced. The organizational entities correspond to branches, divisions, etc., in the real world.

The Operating structure also records the accounting system of the business by showing where in the structure various types of accounting information are recorded and can be made available to the statistical agency. This is accomplished through the use of a series of special flags that indicate the types of accounting data that each Operating entity is able to report (i.e., operating profit, principal inputs, revenues, salaries and wages, and inventories). These flags are an important part of the process to delineate the Statistical structure. The value of these flags is determined through the business profiling process where Statistics Canada contacts a business in order to determine the Legal and Operating structures as well as the accounting system. The data attributes for a given Operating entity are either the sum of the those attributes in the subordinate Operating entities, or provided directly by the business for that particular Operating entity as part of the profiling process.

The Statistical structure for a given business is always made up of four levels; the Enterprise, Companies, Establishments, and Locations. The Enterprise represents the entire business and is associated with the top Organizational entity in the Operating structure. Companies are associated with the lowest level organizational entities that can report operating profits. Establishments are associated with production entities that can report principal inputs, revenues, and salaries and wages. In addition, Statistical Establishments do not cross provincial boundaries. Locations are associated with production entities that are only able to report number of employees. Locations represent a single physical location.

1.2 The Automated Statistical Structure Generator

The four level Statistical structure that is used at Statistics Canada, along with the complex temporal organization of the Business Register database has made the task of generating Statistical structures too difficult to do manually. The automated Statistical structure generator performs the tedious manual work that would otherwise be necessary to delineate, classify and assign data attributes to Statistical entities. Manual intervention is still possible when necessary.

There are several steps in the process used to generate Statistical structures:

1. Modification of the Operating structure including recording the accounting capabilities of each Operating entity in the structure using an on-line interactive dialog.
2. Delineation of the Statistical structure based on the accounting capabilities found in the Operating structure.
3. Classification--the automatic calculation and assignment of activity and geographic codes for the fiscal period of the business.
4. Assignment of data attributes (financial and employment data) from the Operating entities linked to each of the Statistical entities.
5. Updating the Business Register with the new structure. This process takes into consideration the temporal aspect of the database as described above. Continuity of Statistical entities is maintained by minimizing births and deaths, thus providing surveys with a more stable universe for sampling.

2. *Delineating the Statistical Structure*

2.1 Assigning Accounting Capabilities

Businesses provide Statistics Canada with accounting capabilities for each of their component entities as well as the activities that each location is engaged in during the fiscal year covered by a profile. The Business Register is later updated with this information through a series of on-line interactive dialogs called Events. These Events are an approximation of the types of transaction that a business would engage in during the business year, such as amalgamations, acquisitions, reorganizations, selling, buying, opening and closing locations, etc. These Events contain edits which ensure that the information used to update a structure is valid given all of the inter-relationships within the business. This part of the updating process is usually done manually by a human operator because of the complexity of the task and the necessity in most cases to have a familiarity with the business being updated. These events record on the Operating and Legal structures the changes that happened to the business in the real world.

2.2 Associating Operating Entities with the Four Statistical Levels

Once the events to update the Operating structure have been made by the operator and all of the record and structure edits are complete, the operator invokes an automated system called the "flag setter." The flag setter analyses the Operating structure and attempts to determine which entities correspond to each of the four levels in the Statistical structure (Enterprise, Company, Establishment, and Location). This is done by examining the accounting capabilities assigned as described above. The result of this analysis is that each Operating entity is assigned a reporting capability (i.e., an Operating entity corresponding to a Statistical Establishment has "Establishment reporting capability").

2.2.1 Establishment "Roll Ups"

When a given Operating entity has subordinates that do not all have the same reporting capability the flag setter performs a reporting capability "roll up." If, for example, an Operating entity has three subordinates where two of them have establishment reporting capability but the other does not, the flag setter would set a flag on that Operating entity to indicate that some, but not all, subordinate entities have establishment reporting capability. This flag later informs the

Statistical structure generator to generate only one Statistical Establishment for this portion of the Operating structure. This process is referred to as an establishment "roll up" since the two potential subordinate Establishments are "rolled up" into the Operating entity above them (Figure 1 below shows a "roll up"). The flag setter handles a similar situation at the Statistical Company level in the same way.

2.2.2 Pseudo Establishment Forcing Through Operator Intervention

Once the flag setter has completed associating the Operating entities with the levels in the Statistical structure, the operator is provided with a count of the number of Companies and Establishments there would be should the Statistical structure generator be invoked. The operators are trained to look for a significant change in the numbers of Companies or Establishments. Such a change is usually due to errors in updating or profiling, which can be corrected by backing out the incorrect updates and then re-applying them.

In rare circumstances, when the correct data would cause a roll-up, the operator is able to manually force the creation of an Establishment. This is done by setting a special flag that tells the flag setter to associate that Operating entity with a Statistical Establishment even though it does not have the capability of reporting establishment data. This forced Establishment creation is done to meet the desirable objective of having establishment capabilities as low as possible, allowing a more detailed delineation of the Statistical structure, and a more accurate assignment of survey feedback data. Figure 2 on the next page shows the effect of this forcing on the original Statistical structure (Figure 1).

2.2.3 Ancillary Establishments

An Ancillary Establishment is an Operating entity at the same level in the Operating structure as entities with establishment reporting capability, which performs a support role to those Establishments, such as a repair shop or warehouse, and which does not itself have establishment reporting capability. These ancillary entities are normally identified automatically by the flag setter based on their activity codes but can be manually forced if necessary. A similar process occurs at the Statistical Company level.

2.3 Delineating the Statistical Structure

Once the operator is satisfied that the predicted number of Companies and Establishments is correct, the automatic Statistical structure generator is invoked. This processor uses the flags set by the flag setter to generate a Statistical structure in memory that corresponds to the Operating structure. A special case occurs when an Operating entity flagged with establishment reporting capability has subordinate Operating entities in more than one province. In this case the Statistical structure generator creates one Statistical Establishment for each province and links them to each of the subordinate Operating entities for that province (See Figure 1 below).

In some cases such as a buy-sell event, or an amalgamation, more than one Statistical structure is affected. In this case the automatic Statistical structure generator creates new versions of all of the affected Statistical structures in memory.

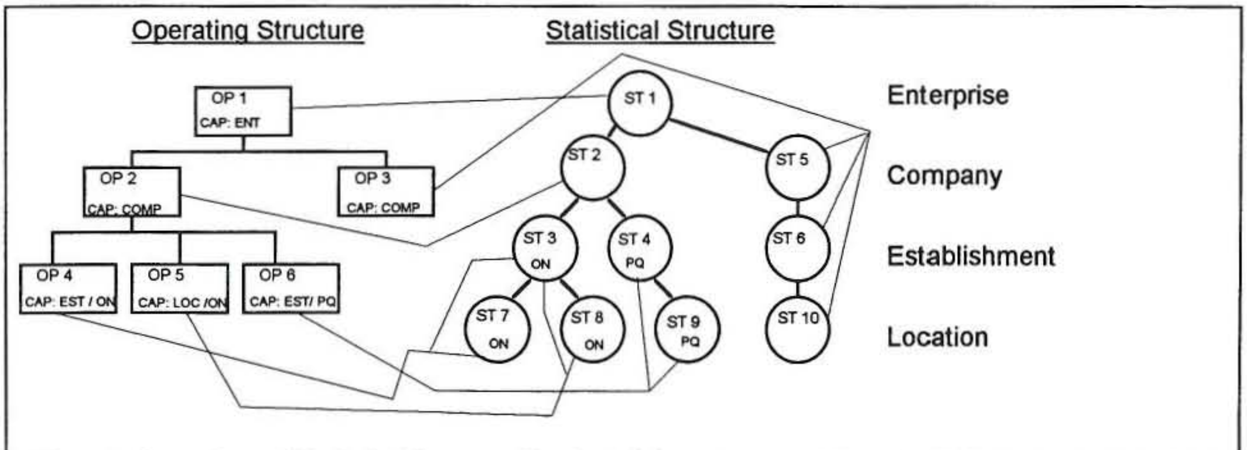


Figure 1: Operating and Statistical Structure Showing Linkages Between Them As Well As a Provincial Split "Roll Up" Establishment (ST 3 and ST 4)

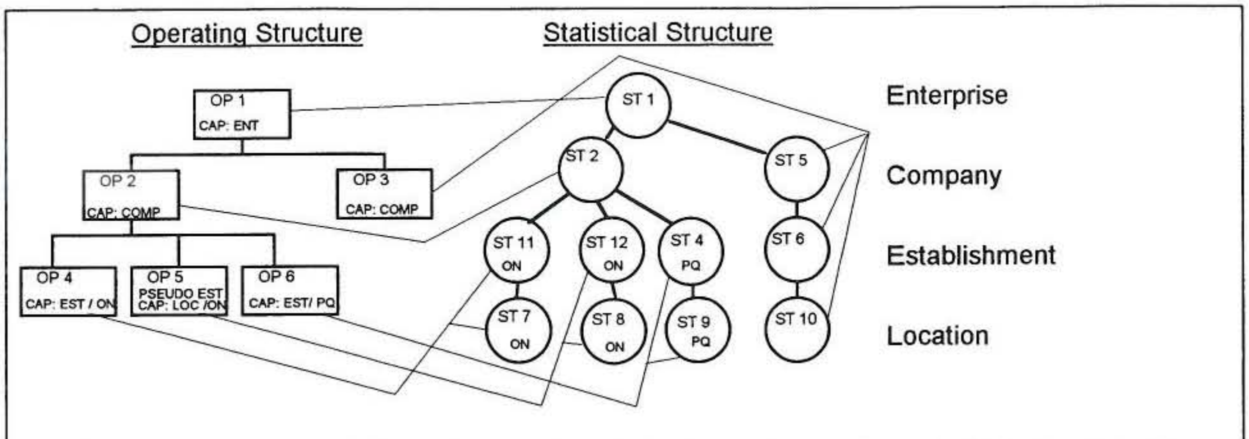


Figure 2: Operating and Statistical Structure Showing Linkages Between Them As Well As a "Pseudo" Establishment (ST 12)

3. *Classification of Statistical Entities and Assignment of Data Attributes*

3.1 Classification

A very important part of the generation of Statistical structures is the classification of the entities that make up those structures. Classification in this case refers to assigning standardized activity and geographic codes. The classification of Statistical entities is based on information found in the Operating structure activity codes and postal codes derived from addresses. The Statistical structure generator takes the variety of data available to it and selects the dominant activities and geographic codes for the Statistical entities. This classification data is assigned to Statistical entities for a complete fiscal year.

3.1.1 Activity Codes

In some cases a given Statistical entity represents operations that perform more than one activity, which may include both support and technical activities. These cases result from either an Operating entity performing more than one activity, or a Statistical entity representing a number of Operating entities. On the Operating entity file, a given production entity may have more than one activity. Each production entity has an activity/commodity table that records all of the significant activities of interest to the Statistical Agency.

The Statistical structure generator takes all of the activities from the activity/commodity table(s) of Operating entities associated with a given Statistical entity and determines which one is the dominant activity. This is done by comparing the results achieved by multiplying the percentage of revenue associated with a particular activity, with a ratio that represents the relative value added to the Gross National Product by revenue from the industry division represented by that activity. For example, if a particular production entity generated 55% of its total revenue of $100,000 from wholesale operations and 45% from retail operations, then the associated Statistical entity would be given a retail SIC. This is because the ratio for retail (.168) when multiplied by the smaller percentage yields a larger result ($100,000 x 0.168 x 45% = $7650) than when the ratio for wholesale (.112) is multiplied by the larger percentage ($100,000 x 0.112 x 55% = $6160) (See table 1 below for the ratios used for each industry).

INDUSTRY	RATIO
Agricultural and Related Industries	.526
Fishing and Trapping Industries	.692
Logging and Forestry Industries	.330
Mining, Quarrying and Oil Wells Ind.	.527
Manufacturing Industries	.289
Construction Industries	.430
Transportation and Storage Industries	.555
Communication and Other Utilities	.657
Wholesale Trade Industries	.112
Retail Trade Industries	.168
Finance and Insurance Industries	.237
Real Estate and Insurance Agents	.557
Business Service Industries	.703
Government Service Industries	N/A
Educational Service Industries	.869
Health and Social Service Industries	.799
Accommodation, Food & Beverage Ind.	.504
Other Service Industries	.576

Table 1: Value Added Ratios by Industry Division

For the lower two levels in the Statistical structure, activities are coded according to the Standard Industrial Classification for Establishments (See the manual *Standard Industrial Classification: 1980* for more details) based on the activities of linked production entities. The upper two levels, Enterprises and Companies, are coded according to the Company Classification System (See the manual *Canadian Standard Industrial Classification for Companies and Enterprises: 1980*) based on the activities of the subordinate entities in the Statistical structure only.

3.1.2 Geographic Codes

Calculation of Geographic codes is done using a geographic database that is accessed with the postal code of the dominant Operating entity corresponding to a given Statistical entity. In the case where there is more than one donor Operating entity to choose from, the operation with the largest revenue and selected dominant activity is chosen as the postal code source. This geographic coding is only done for Statistical Establishments and Locations.

3.2 Assignment of Data Attributes

Economic data stored on Operating entities are assigned directly to the Statistical entity associated with them. Data attributes are summed before assigning them to the associated Statistical entity when there is more than one Operating entity associated with that Statistical entity. This data includes total assets, depreciable assets, revenue, and number of employees.

4. *UPDATING THE DATABASE*

4.1 Rules for Birthing and Deathing Statistical Entities

All of the delineation, classification, and assignment of data attributes to the Statistical structure are done in computer memory with no updates to the Business Register taking place. Once all of the data have been calculated, a comparison is made between the Statistical structure(s) stored in memory, and the one(s) stored on the Business Register. Every effort is made to maintain continuity between the old and new Statistical structures by updating existing entities when possible and birthing and deathing other Statistical entities only when necessary. Sets of rules have been devised in order to handle all cases consistently for each level in the Statistical structure. The rules described below have been simplified somewhat for this paper. A complete description of them is found in Armstrong and Cuthill, 1988.

4.1.1 Rules for Birthing and Deathing Enterprises

The Statistical Enterprise (top level of Statistical Structure) is always linked directly with the top Operating entity. This means that the top Operating entity, and the Statistical Enterprise remain the same for the life of the business.

4.1.2. Rules for Birthing and Deathing Companies

1. If a new Statistical Company is linked to the same Operating entity as a Company in the previous Statistical structure then the old Company entity is updated with the new data.
2. If all the Establishments subordinate to the Statistical Company are births then the Statistical Company is considered a birth.
3. If the above two rules do not handle all of the Company entities in the new Statistical structure then a search is made of the Companies in the old Statistical structure to find one that has donated all of its subordinate Statistical Establishments to the new Company. If one is found, then that old Statistical Company entity is updated with the new data. In all other cases the Company entity in the new structure is considered a birth.
4. All Statistical Companies in the old Statistical structure(s) that no longer have any subordinate Establishments are deathed.

4.1.3 Rules for Birthing and Deathing Establishments

1. If all Statistical Locations subordinate to an Establishment are births, the new Establishment is considered a birth.
2. If an Establishment is linked to the same Operating entity as an Establishment in the previous structure then the old Establishment entity is updated with the new data.
3. If the above two rules do not successfully determine whether a new Establishment is a birth or an update to an existing Establishment, the Locations linked to the new Establishment are classified according to whether they were originally from the same business, or from another business. The previous parents of each of the Locations are analyzed to find a list of potential Establishment donors. The basic rule that is then applied is that if any one Establishment is found that has donated all of its Locations except for those which were deathed or sold then that

Establishment will be updated with the new Establishment data. In all other cases a new Establishment entity is birthed.

4. All Statistical Establishments in the old Statistical structure(s) that no longer have any subordinate Locations are deathed.

4.1.4 Rules for Birthing and Deathing Locations

The Statistical Location level is linked directly to the lowest level entities on the Operating structure (production entities) so that when an Operating "leaf" is deathed or birthed a corresponding Statistical Location (lowest level of Statistical structure) is deathed or birthed.

4.2 The Option of Operator Intervention

All of the births, deaths and updates to the Statistical structure described above are done in a "non-viewable" mode, meaning that they are not released for retrieval by the Business Register users. This allows the operator, and others, to confirm that the set of updates resulted in what was expected. In the case where the updates are judged to be incorrect, the updates can be "backed out" (one at a time in the opposite order to which they were performed) to restore the Operating and Statistical structures to the state they were in before the updates occurred. Only the updates back to the one in error need be "backed out" and the operator assigned to the work only needs to redo the few Events that were "backed out." Once the updates are judged to be correct the information is made viewable for access by surveys and other users. This iterative process allows sufficient human intervention when necessary to handle the difficult situations that can arise.

5. Conclusion

The IP Statistical structure generator has an important part in producing timely Statistical structures for surveys and other users of economic data at Statistics Canada. The majority of the tedious work that used to be involved in delineating and creating Statistical structures on the previous Business Register has been eliminated. Today, generation of a Statistical structure involves merely pushing a function key following the on-line interactive dialog that is used to update the Operating and Legal structures. In most cases the Statistical structure is generated within a few seconds and can be browsed on-line to ensure that it is as expected.

The Statistics Canada Business Register has been designed in order to provide various kinds of economic data for a wide range of users The Statistical structures are used to provide that data in standardized form. This has resulted in an innovative, flexible design that is also necessarily quite complex. The Statistical structure generator described in this paper is one of the most complex sub-systems in the Business Register. Two of the challenges encountered with this design have been to convince skeptics that the automatic process would produce as good a result as a manual process, and to train new users, operators, and systems maintainers to take full advantage of all of the capabilities of the system. These challenges and others have largely been met as the Business Register continues to provide high quality economic data to an ever-increasing population of users.

6. References

Armstrong, Gerry (1988). "An Overview of the Generation of Statistical Entities." Statistics Canada Internal Presentation, December 6, 1988.

Armstrong, Gerry, and Cuthill, Ian (1988). "Detailed Specifications for the Automatic Statistical Generator." Working Paper of the Business Survey Redesign Project. Statistics Canada. May 1988.

Canadian Standard Industrial Classification for Companies and Enterprises: 1980. Statistics Canada, Standards Division, Ottawa. ON. 1986. (Catalog 12-570E.)

Colledge, Michael and Armstrong, Gerry (1989). "Statistical Units, Births and Deaths At Statistics Canada After the Business Survey Redesign." Presented at the Third International Roundtable on Business Survey Frames, Auckland, New Zealand, November 1988.

Cuthill, Ian M. (1989). "The Statistics Canada Business Register." *Proceedings of the U.S. Bureau of the Census 1989 Annual Research Conference*, 69-86.

Hamm, Carole. "An Overview of the Business Register: A Computer Assisted Instruction Course." Version 1.0. Statistics Canada, Business Register Division, September 1992.

Standard Industrial Classification: 1980. Statistics Canada, Standards Division. Ottawa, ON. 1980. (Catalog 12-501E.)

A NEW SWEDISH BUSINESS REGISTER COVERING A CALENDAR YEAR AND EXAMPLES OF ITS USE FOR ESTIMATION

Eva Elvers, Statistics Sweden
SCB, S-115 81 Stockholm, Sweden

KEY WORDS: Register, frame, coverage, estimation

Abstract

The Swedish business register is updated regularly and provides as recent information as possible on enterprises and local units. The target population of a survey should refer to the same period as the statistics do, rather than to the situation as given by the sampling frame. Statistics Sweden has constructed a new type of business register that covers a calendar year. This register and its use for estimation in the industrial survey and for comparisons of investments estimated by three surveys are described.

1. Introduction

The population of enterprises and local units changes, which causes problems with coverage and classifications. Statistics Sweden has initiated improvements of its registers in this respect, first by constructing business registers covering a calendar year.

The traditional Swedish business register (BR) is described in this section and the new register in section 2. The first two uses of the new register are presented in sections 3 and 4. Section 5 concludes.

The BR of Statistics Sweden is called the Central Register of Enterprises and Local Units. Central pieces of information are, of course, kind of activity (industry), size, and geographic location. Kind of activity or industry here means the SIC code, the Swedish version of the United Nations International Standard Industrial Classification (ISIC). The size measure is the number of employees. The BR obtains information from several sources.

There are two main levels in the BR. There is the enterprise level consisting of legal units and physical persons, who among other things are registered for value added tax (VAT). There is the local unit level, where the address is an important piece of information.

The BR obtains information on births and deaths from the National Tax Board every second week. The number of employees is updated once a year through the Tax Payroll (PAYE) and through a special questionnaire to multiple-location enterprises. There is also information from the surveys of Statistics Sweden on changes discovered during the data collection. The annual Industrial Survey (IS) is an important source for the SIC code.

There is a modified version of the BR, called the Statistical Register (SR), which is used as frame for the annual and sub-annual business surveys. Some units consist of a set of legal units. These units are the smallest ones for which balance sheet and profit and loss data can be obtained. They are essential to the Financial Accounts Survey (FAS), and they are used by all surveys for the sake of comparability. There are about 60 large such statistical units consisting of more than 400 enterprises. Most samples are drawn in the so-called SAMU system, Ohlsson (1992, 1993). The emphasis here is on annual surveys.

2. BRs: events, a situation, and a time period

2.1. A situation register

The BR is based on several sources. Information on events - births and deaths - is combined with the previous version of the register to create a new version, the most recent description of enterprises and local units. The BR is a snapshot of the changing population, a *situation register* describing enterprises and local units at a certain point in time. It is more correct to say that the BR describes several situations, since some variables are updated more frequently than others.

In the SAMU system samples are normally drawn in November the year that the annual surveys are to investigate, year t. All surveys use industry (the SIC code) for stratification. Most surveys also stratify by size, and the size measure is mostly the number of employees.

In November year t we can expect the SR to describe the situation in the end of September as to active enterprises and local units. Deaths before that time, t-deaths, and births after that time, t-births, are not in the frame. The number of employees refers to the spring for multiple-location enterprises (questionnaires) and to December year (t-1) for single-location enterprises (PAYE information). Single-location enterprises born year t have 0 employees in the BR. Hence surveys that cover enterprises with a minimum number of employees only do not cover births year t.

2.2. A register covering a time period

Business statistics refer to a time period. Statistics from annual surveys usually refer to a calendar year. To improve the coverage of the statistics it is reasonable to construct a register that covers the same *time period*. Statistics Sweden has recently started such work.

We call such a register an ÅR. The Swedish letter Å denotes a year, a *calendar year*. The ÅR is an extension of the SR comprising all enterprises and local units that were active during a year, the whole year or part of it. So far, only the enterprise level is completed, whereas the local unit level is still under development.

The ÅR describing year t, $ÅR_t$, is now completed in the beginning of year (t+2), about 14 months after the end of the period referred to. By then the IS year t is finished, and the information on the large statistical units has been improved. There is a special register for the latter, UR.

The SIC code in $ÅR_t$ is taken from

(i) UR_t,

(ii) IS_t,

(iii) SR_t for enterprises outside IS_t,

(iv) SR_{t+1} for t-births, and

(v) SR_{t-1} for t-deaths.

There are some more rules to take care of reorganizations of enterprises.

The size measure, i.e. the number of employees, is to be a measure referring to the whole year. The sources used are

(i) UR_t,

(ii) the BR questionnaire for multiple-location units, which means the number of employees in the SR_t,

(iii) the average of the number of employees according to the SR_t and the SR_{t+1} for single-location units active throughout the whole year, which means the average of the numbers in December the years (t-1) and t.

For singe-location units active part of the year, the existing number is taken.

There are about 495 000 enterprises in the 1991 SR, nearly 560 000 active enterprises in the 1991 ÅR, and somewhat more than 560 000 inactive enterprises in a connected register. The difference between the ÅR and the SR is due to 53 000 t-deaths and 12 000 t-births in 1991. The ÅR contains starting and closing dates for all enterprises that have not been active throughout the whole year.

For an annual survey with the population restricted to units with a minimum number of employees, the difference in coverage when using the information in the $ÅR_t$ instead of that in the SR_t will be larger than the difference for the whole population. There are enterprises in the SR_t with 0 employees due to lack of information, and enterprises increase and decrease their number of employees so that they become in-scope and out-of-scope enterprises, respectively.

In 1992 ÅRs of enterprises were derived for the five years 1986-90. The choice of as many as five years was made to enable time series, and all these ÅRs have been used for the IS. In 1993 a first ÅR of local units is developed. Moreover, the enterprise ÅR is improved in a few respects.

2.3. Population and domains of estimation

For most surveys, the population and the domains of estimation are the same as those given by the frame, i.e. the SR_t situation as to active enterprises and local units and classifications. Deaths are exceptions, as well as a small number of enterprises showing that they are out-of-scope due to a wrong classification. Many surveys do not collect information on the basic variables SIC code and size.

This means that statistics for year t refer to an old population with old classifications. There are at least two reasons to keep to old classifications. If new classifications are obtained for the sample only, the variance of the estimator of a domain total may be considerable. The (squared) bias due to old classifications may be comparatively small, and the old classifications may be preferred for that reason. A second reason is to ensure consistency between surveys, important for the National Accounts, among others.

The ÅR provides us with populations and classifications referring to year t. The new classifications are known both for each sample and for the whole population. Hence, when the domains of estimation are kind of activity (SIC code) we know the population size of each domain. Moreover, we can use the new size measure for post stratification. Hence, we will overcome both reasons for keeping the old classifications.

3. Utilizing the ÅR for a renewed Industrial Survey

3.1. The Industrial Survey

The target population of the IS was changed in the 1990 survey. Now all enterprises classified as 'mining and quarrying or manufacturing' (manufacturing for short below) and with at least 10 employees are included, and also manufacturing establishments with at least 10 employees within non-manufacturing enterprises.

Questionnaires are sent to all enterprises, and all establishments with at least 5 employees are to answer a separate establishment questionnaire. The frame year t is based on information from the SR_t and the IS_{t-1}. Responses in the IS_t determines SIC code and size that year. There are nearly 10 000 establishments in the IS. Many an establishment coincides with a local unit in the BR.

The IS contains many variables measuring

(i) operating income - total and different kinds,

(ii) operating expenses - total and different kinds of expenses, e.g. compensation of employees, raw materials, and hired transports,

(iii) stocks,

(iv) investments,

(v) numbers of salaried employees and wage-earners,

(vi) quantities and purchased values of fuels and electric energy consumed, and

(vii) quantities and values for commodities produced.

3.2. Utilizing the ÅR, the FAS, and administrative data

When discussing the IS and estimation methods for the new population, two main decisions were taken,

(i) to use the ÅR to determine the population on the enterprise level and

(ii) to utilize administrative data for nonresponse and undercoverage estimation.

The IS is a source of information of the ÅR, and all establishments are surveyed. Hence, the effect of using of the ÅR will be less pronounced than for most other surveys. The population is mainly extended by enterprises not being in-scope according to the SR. A set of nonresponding enterprises with fewer than 10 employees according to the ÅR is excluded from the population.

The FAS provides information on total operating income, total operating expenses, the total of salaries and wages, employer´s contributions to social security, and investments on the enterprise (statistical unit) level. It covers all units with at least 20 employees (and a sample for smaller units; there is some nonresponse in the FAS too). The PAYE provides information on the total of salaries and wages (on a local unit level with a grouping that partly differs from that of the BR, and also from that of the IS), and the VAT register provides information on the turnover on the enterprise level. When registers are matched care is needed, e. g. in handling reorganizations with old and new identification numbers.

The ÅR and the administrative data are used not only for the regular IS (as described in the next section) but also to obtain statistics on some basic variables for industrial enterprises with fewer than 10 employees. There are about 40 000 such enterprises in 1990, and they are not surveyed at all. They are now included in the overall estimates of five variables. They are treated as single-establishment enterprises.

3.3. Estimation procedure for nonresponse and undercoverage

The new estimation procedure uses several sources of information. Values are imputed on the establishment level for both nonresponse and undercoverage with the aim of obtaining good quality of the statistics by SIC code.

The first step in the imputation procedure is to use information (values of variables) in the FAS if available and otherwise from the PAYE and/or the VAT register. For a multiple-location enterprise, some extra, mostly manual work is needed to split the values into establishments as far as possible. The second step in the imputation procedure is to estimate the remaining variables given the values from the first step and the IS the present and the previous years.

On an intuitive basis it seems reasonable that many establishments this year will use the same fuels, have approximately the same proportions of salaries and wages et cetera, as they did last year. If so, subtotals can be estimated by splitting totals in the proportions of the response the previous year. When estimating quantities changes in prices of energy and commodities and in compensation of employees have to be considered. A second basis for estimation purposes is the standard assumption of homogeneity within groups of kind of activity and size.

After having tested these ideas on establishments responding two adjacent years we use the previous-year-method when there is a response unless the establishment has changed considerably in some respect (considering SIC code and the variables imputed in the first step).

About 20 % of the establishments are imputations, but nonresponse is more frequent among small establishments than among large ones. Measured relative to totals, about 10 % of the value is imputed for a few central variables.

4. Comparing statistics on investments

Our second use of the ÅR has been as a basis for comparison of statistics on investments. There are three surveys providing information on investments with somewhat different aims. Differences have been observed; differences so great that they ought to be analyzed. Comparisons made earlier were not very conclusive. The present study based on the ÅR is not finished yet, but some experiences have been gained.

The three surveys are the IS, the FAS, and the Investment Survey (InvS). The InvS is a sub-annual survey covering enterprises with at least 20 employees. Investment plans on different time horizons are important variables. In the February survey year (t+1) investments made in year t are reported. The sample was drawn in SAMU year (t-1), so the frame is one year older than that of the FAS. A stratification according to SIC code and size is used.

All three surveys have been matched to the ÅR on the enterprise (statistical unit) level for manufacturing units with at least 20 employees. For enterprises that have been reorganized there are considerable differences between the surveys in their handling of identification numbers. This makes comparisons on the unit level more difficult. Comparisons of the statistics are, of course, more important.

The most pronounced difference in the comparisons by SIC code and size is that the InvS has a low total for small units, which are in take-some-strata. This came as a surprise to the InvS which has been more concerned about the large enterprises and their reorganizations. Less than 20 % of the investments are made by enterprises with 20-99 employees. Comparisons of the respondents in such strata show that enterprises which have diminished to less than 20 employees (out-of-scope-units) have much lower values than enterprises which remain in-scope units. FAS data indicate that investment differences between units in-scope already year (t-1) and units entering the population year t are relatively small. This will be studied further.

The ÅR has shown us the problem and its importance, but the ÅR itself is not a solution of the InvS problem of an old population, since the ÅR is available too late for this survey. We will consider using other register data, and we will investigate the possibility of a supplementary sample as well as other estimation methods for undercoverage in this survey.

5. Conclusions

So far the ÅR has had two uses. It has contributed in several respects to the improvements of the IS, and it has provided useful information on the statistics on investments.

Next some further survey will utilize the ÅR. We then expect improvements in the population and the domains of estimation. We will study methods of nonresponse estimation.

ÅRs are being developed not only for enterprises but also for establishments. In the IS we have observed through a comparison with the PAYE, that further information on manufacturing establishments within non-manufacturing enterprises and vice versa would be valuable.

6. References

Ohlsson, E. (1992). SAMU. The system for Co-ordination of Samples from the Business Register at Statistics Sweden. R&D Report 1992:8, Statistics Sweden.

Ohlsson, E. (1993). Co-ordination of several samples using permanent random numbers. To appear in the ICES monograph.

BENCHMARKING OF SMALL AREA ESTIMATORS

H.J. Mantel, A.C. Singh, and M. Bureau, Statistics Canada
H.J. Mantel, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario K1A 0T6

KEY WORDS: Direct and indirect estimators, composite estimators, bias and MSE

1. Introduction

There exists a considerable body of research on small area estimation using cross-sectional survey data in conjunction with supplementary data obtained from census and administrative sources. A good collection of papers on this topic can be found in Platek, Rao, Särndal and Singh (1987). For large areas (or domains) direct estimators (*i.e.* estimators based only on sample data from the area of interest) are often used; however, indirect estimators, in which strength is borrowed from similar areas via a model containing auxiliary variables from the supplementary data, are often used for small areas. For repeated surveys it may also be beneficial to borrow strength over time; see Pfeffermann and Burck (1990) and Singh and Mantel (1991). Direct small area estimators, though (approximately) unbiased, are not reliable because of high variance. Indirect small area estimators are more reliable, though they may be somewhat biased.

A common problem in the application of small area techniques is that the individual small area estimates within a larger area do not add up to the direct estimator for the larger area. This problem can be resolved by benchmarking of the small area estimators with respect to the direct estimator for the larger area. This is desirable for at least three reasons: (*i*) the usual direct estimator for the larger area is approximately unbiased, whereas the aggregated small area estimators may be substantially biased, (*ii*) benchmarking gives rise to some robustification in that the average of the benchmarked small area estimators has good bias and variance properties, (*iii*) there will be internal consistency between published estimates for the larger area and the total of estimates of the individual small areas within it.

Three methods for benchmarking are proposed in the literature: (*i*) Battese, Harter and Fuller (1988) distribute the difference between the direct large area estimator and the sum of the small area estimators in proportion to the mean squared error (MSE) of each small area estimator. (*ii*) Pfeffermann and Barnard (1991) distribute the difference "optimally" using the full MSE matrix of the small area estimators. This method has an advantage for time series methods in that it can be built in as part of the Kalman filter algorithm (giving as a byproduct an estimate of the MSE matrix of the benchmarked estimators); see Pfeffermann and Burck (1990). (*iii*) Rao and Choudhry (1993) distribute the difference in proportion to the small area estimates, *i.e.* a simple ratio (or raking) adjustment is made.

In this paper we perform an empirical study using a synthetic population based on data from Statistics Canada's Survey of Employment, Payroll and Hours (SEPH) to compare the effect of benchmarking on various small area estimators. In particular, we compare, in a repeated sampling framework, the loss in efficiency due to benchmarking to the gain in efficiency due to "borrowing strength". Two types of indirect small area estimators are synthetic (in which small areas are assumed to be like a larger area) and composite (convex combinations of direct and synthetic estimators). For small area estimation we consider three types of composite estimators where the weights for the convex combination can be either (*i*) optimal (*i.e.* based on a correctly specified model), (*ii*) pseudo-optimal (*i.e.* based on an incorrect model), or (*iii*) based on some other working convention such as the one for sample size dependent weights.

2. Domain Estimation Methods

Let the vector of small area population totals, Y_a, $a=1, \dots A$, be denoted by $\underset{\sim}{Y}$. Here we define briefly some well known small area estimators which we will use in our simulation study. Rao (1986), Särndal and Hidiroglou (1989) and Pfeffermann and Burck (1990) also contain a good survey of various small area estimators.

2.1. Direct Estimators

2.1.1 Expansion estimator

This method of estimation is defined by $\mathrm{EXP}_a = \Sigma_{i \in s_a} w_i y_i$ where s_a is the portion of the sample falling in small area a, and w_i is the survey weight for unit i. For stratified simple random sampling, which we use in our simulation study, we have

$$\mathrm{EXP}_a = \Sigma_k (N_k/n_k) \Sigma_{i \in s_{ka}} y_i \ , \qquad (2.1)$$

where s_{ka} denotes the set of n_{ka} sample units falling in the small area a and stratum k and n_k, N_k denote respectively the sample and population sizes for the kth stratum. The above estimator is often unreliable because the random sample size n_{ka} may be small in expectation and could have high variability. Conditional on the realized sample size n_{ka}, $\mathbf{EXP}_a$ is biased; however, unconditionally, it is unbiased for Y_a.

2.1.2 Separate ratio estimator

If X_{la}, the small area total of a suitable covariate, is known for some post-strata indexed by l, then the efficiency of the estimator $\mathbf{EXP}_a$ could be improved upon by exploiting this knowledge. We define

$$\mathrm{SRAT}_a = \Sigma_l X_{la} \hat{Y}_{\exp,la} / \hat{X}_{\exp,la}\,, \qquad (2.2)$$

where $\hat{Y}_{\exp,la}$ is the expansion estimator for the total of y in small area a by post-stratum l. In our simulation study later we take the post-strata to be the intersection of design strata with small areas. When the covariate x is a constant then the estimator, also called post-stratified and denoted by $\mathbf{POST}_a$, is both conditionally and unconditionally unbiased; however, $\mathbf{SRAT}_a$ would generally be slightly biased. These estimators may also be not sufficiently reliable because of the possibility of n_{ka}'s being small in expectation. If $\hat{X}_{\exp,la} = 0$, the above estimators are not defined. In practice, some ad hoc value such as 0 is often chosen for $\hat{Y}_{\exp,la}/\hat{X}_{\exp,la}$ when $\hat{X}_{\exp,la} = 0$. In the simulation study presented in this paper, we set $\hat{Y}_{\exp,la}/\hat{X}_{\exp,la} = \hat{Y}_{\exp,l}/\hat{X}_{\exp,l}$ whenever $\hat{X}_{\exp,la} = 0$.

2.1.3 Combined ratio estimator

An alternative to the separate ratio estimator is the combined ratio estimator,

$$\mathbf{CRAT}_a = X_a \mathrm{EXP}_a / \hat{X}_{\exp,a} \qquad (2.3)$$

When the covariate x_i is a constant then the estimator will be denoted by $\mathbf{HAJEK}_a$. $\mathbf{CRAT}_a$ would generally be slightly biased. If $\hat{X}_{\exp,a} = 0$ then the above estimators are not defined. In practice, some ad hoc value such as 0 is often chosen for $\mathrm{EXP}_a / \hat{X}_{\exp,a}$ when $\hat{X}_{\exp,a} = 0$. In our simulation study presented later, we set $\mathrm{EXP}_a / \hat{X}_{\exp,a} = \hat{Y}_{\exp} / \hat{X}_{\exp}$ whenever $\hat{X}_{\exp,a} = 0$.

2.1.4 Generalized regression estimator (GREG)

In this method a linear regression model is assumed to relate the individual level variate values y_i to a vector of covariates x_i. These covariates would need to be known for each sampled unit and domain totals would also be required. The sample data can be used to estimate the regression parameter and a synthetic estimator of the domain totals is then constructed. However, there may be some local lack of fit of the global regression model and this is accounted for by a direct estimate of the domain sum of residuals from the regression. The estimator is

$$\mathrm{GREG}_a = x_a^T \hat{\beta} + N_a \bar{e}_a \qquad (2.4)$$

where $\hat{\beta} = (\Sigma_s (x_i x_i^T)/(\nu_i \pi_i))^{-1} (\Sigma_s (x_i y_i)/(\nu_i \pi_i))$, $\bar{e}_a = \hat{e}_{\exp,a} / \hat{N}_{\exp,a}$, $e_i = y_i - x_i^T \hat{\beta}$, x_a is the domain a total of the covariate vectors x_i, ν_i are pre-specified regression weights and π_i is the survey weight for unit i. This version of generalized regression estimation, with a synthetic $\hat{\beta}$, was proposed by Särndal and Hidiroglou (1989). When the sample size in domain a is 0 we take $\bar{e}_a = 0$. $\bar{e}_a$ would be relatively stable when the regression model accounts for a large proportion of the variability in y.

2.2 Composite Estimators

2.2.1 Sample size dependent estimator

If the observed sample size in small area a is small then we may consider a convex combination of a direct estimator and a synthetic estimator (*e.g.* $x_a^T \hat{\beta}$ of (2.4)). Using sample size dependent weights, we have

$$\mathrm{SSD}_a = (1 - \lambda_a)\hat{Y}_{\mathrm{syn},a} + \lambda_a \hat{Y}_{\mathrm{dir},a} \qquad (2.5)$$

where $\lambda_a = 1$ if $\hat{N}_{\exp,a} \geq N_a$ and $\lambda_a = (\hat{N}_{\exp,a}/N_a)^d$ otherwise, and d is assigned some suitable value such as 1 or 2.

2.2.2 Empirical best linear unbiased estimator (EBLUP)

An alternative to sample size dependent smoothing of small area estimators is to use the empirical Bayes approach of Fay and Herriot (1979) or the more general best linear unbiased predictor (BLUP) approach (see *e.g.* Battese, Harter, and Fuller (1988), and Pfeffermann and Barnard (1991)). It is assumed that $\underset{\sim}{Y} = F\underset{\sim}{\alpha} + \underset{\sim}{\nu}$ where the ν_as are small area effects and F is a matrix of regressors. The model for the small area estimators is then $\hat{\underset{\sim}{Y}}_{\mathrm{dir}} = F\underset{\sim}{\alpha} + \underset{\sim}{\nu} + \underset{\sim}{\varepsilon}$ where ε_a is an observation error term. The BLUP under this model is

$$\mathrm{BLUP} = \Lambda \hat{\underset{\sim}{Y}}_{\mathrm{dir}} + (I - \Lambda) F \hat{\underset{\sim}{\alpha}} \qquad (2.6)$$

where $\Lambda = V(V+W)^{-1}$, V and W are, respectively, the MSE matrices of $\underset{\sim}{\hat{Y}}_{\text{dir}}$ and $F\underset{\sim}{\hat{\alpha}}$, and $\underset{\sim}{\hat{\alpha}}$ is the generalized least squares estimate of $\underset{\sim}{\alpha}$. The mean squared error of **BLUP** is given by $V - V(V+W)^{-1}V$. The variance components V and W would need to be estimated, a survey based estimate would be used for V and then W would be estimated conditional on the estimated V using Henderson's method; more details are given in Section 3. When V and W are replaced by estimates the resulting estimator is termed empirical **BLUP** or **EBLUP**. When the model for the direct estimators is correctly specified the resulting estimator would be called optimal, otherwise it would be called pseudo-optimal.

2.3 Benchmarking

It is sometimes desirable that small domain estimators should add up to direct estimators for certain larger domains containing them. One simple possibility, presented by Choudhry and Rao (1992) is to make a ratio adjustments within each larger area. We will indicate this ratio adjusted constrained estimator by the prefix **CR_** (*e.g.* **CR_EBLUP** for the adjusted **EBLUP**). A second approach, following Pfeffermann and Barnard (1991), and which we will indicate by the prefix **CD_**, is based on the MSE (dispersion) matrix for the small area estimators. If the constraint is expressed as $L^T\underset{\sim}{Y} = \underset{\sim}{c}$, with $\underset{\sim}{c}$ a fixed, known constant, then the minimum MSE linear unbiased estimator is

$$\underset{\sim}{\hat{Y}} + \Gamma L(L^T \Gamma L)^{-1}(\underset{\sim}{c} - L^T\underset{\sim}{\hat{Y}}) \qquad (2.7)$$

where $\Gamma = \mathbf{MSE}(\underset{\sim}{\hat{Y}})$. The third approach, suggested by Battese, Harter and Fuller (1988), and denoted by the prefix **CV_**, is given by (2.7) with the off diagonal elements of Γ set to zero.

3. Simulation Study

The methods described in Section 2 were compared empirically by means of a Monte Carlo simulation from a synthetic pseudo-population based on data from Statistics Canada's Survey of Employment, Payroll and Hours (SEPH). The SEPH sample is currently stratified by 1980 three digit standard industrial classification (SIC3) within province and four size classes; however, under a proposed redesign of the survey the sample will no longer be controlled at the SIC3 level, but rather at some aggregation of SIC3s such as SIC2. An objective of the research reported in this paper is to investigate methods for estimation at the SIC3 by province level after the redesign. Because the sample will no longer be controlled at the SIC3 level this is a domain estimation problem. Larger establishments, and those with a complex structure, are subject to higher sampling rates so that direct estimates at the SIC3 level are satisfactory. However, for smaller establishments (size strata 1 and 2) of simple structure (in what is called the non-integrated portion of the frame, NIP) small domain estimation techniques could be necessary for production of SIC3 by province level estimates. A covariate which can be used for these units is PD7 data which records monthly income tax payroll deductions submitted to Revenue Canada.

To construct the pseudo-population used in our study, we took sample data from the province of Ontario for SIC1=3 (industrial manufacturing and products) and the NIP portion of size classes 1 and 2. Variables included were the SIC3 code, the number of employees, the 3 month average PD7 remittance, the size classification, and the survey weight. We used this data to fit the model

$$y_{ijk} = x_{ijk}(\beta + \nu_i + \xi_{ij} + \varepsilon_{ijk})$$

where y_{ijk} is the number of employees for the kth unit in the jth SIC3 in the ith SIC2, x is the 3 month average PD7 remittance plus 500, β is fixed, and ν, ξ, and ε are independent random components. Using the survey weights as replicate weights, we expanded the pseudo-population, which had 995 distinct units, to 24,074 units. The pseudo-population contained 42 SIC3s (small areas) in 9 SIC2s (*e.g.* fabricated metal products industries, non-metallic mineral products industries). The small area population sizes varied from 26 to 14,236 units. We generated new numbers of employees from the fitted model, except that the estimated variance components were scaled down to reduce the problem of zeros in the data. We simulated sampling from this pseudo-population using stratified simple random sampling by size class and SIC2. The sample size for each stratum was taken to match the total SIC2 by size class in the SEPH sample, though the sampling fractions at the SIC3 level would differ from the SEPH sample. The expected sample size within small areas varied from 1.10 to 142.16 and averaged 23.69.

3.1 Estimation methods used in the study

All of the general estimation methods described in Section 2 were included in the study, with some particular features as described here. Since SIC3s are entirely contained in the corresponding SIC2, each

SIC3 crossed at most two of the design strata corresponding to the two size strata within the SIC2.

The estimators **EXP**, **POST**, **SRAT**, **HAJEK** and **POST** are exactly as described in Section 2.

The remaining unbenchmarked estimators were applied separately within each size stratum and all further discussion of them in this subsection should be taken as being within size classes.

For the GREG estimator, the parameter β has two components, one corresponding to a constant term, and the second corresponding to $\boldsymbol{x}_i$, the PD7 remittance plus 1000 (to avoid the problem of 0 remittances). All sample data within the SIC1 were used in the estimation of β and v_i was taken to be $\boldsymbol{x}_i$.

Two sample size dependent estimators are considered, both with $\boldsymbol{d}=2$ and with the synthetic part being $x_a^T\hat{\beta}$, where $\hat{\beta}$ is defined as in Section 2.4. The first, which we denote by **SSD**, has the estimator **POST** as the direct part; the second, denoted by **SSD***, has **GREG** as the direct part. The estimator **SSD*** was proposed by Särndal and Hidiroglou (1989).

There are four versions of the **EBLUP** estimator considered, based on two direct estimators, **POST** and **GREG**, and two different models. Both models take the matrix $\boldsymbol{F}$ as including a column of 1's and a column of $\boldsymbol{x}_a$'s, the small area totals of $\boldsymbol{x}_i$, where $\boldsymbol{x}_i$ is as for the **GREG** estimator. They differ in how they model the small area effects, ν_a. In the first we model them as $\nu_a = x_a^{\gamma/2}(\nu_k + \xi_a)$ where x_a is the domain $\boldsymbol{a}$ total of $\boldsymbol{x}_i$, ν_k is a random effect that is common to all SIC3s within the same SIC2 $\boldsymbol{k}$, and ξ_a is a random effect for SIC3 $\boldsymbol{a}$. It was assumed that $\nu_k \sim (0,\sigma_\nu^2)$, $\xi_a \sim (0,\sigma_\xi^2)$, and all random effects and the observation errors ε_a are independent. The standard variance estimator for simple random sampling without replacement was used for the entries of $\boldsymbol{V}$ (which is diagonal, estimation of β for **GREG** was ignored in estimation of $\boldsymbol{V}$). When the observed sample size in an SIC3 was 1 a synthetic estimator of the design variance based on data from the corresponding SIC2 was used, and when the observed sample size was 0 the MSE was taken as infinity. Taking the estimated $\boldsymbol{V}$ as the true value, the variance components σ_ν^2 and σ_ξ^2 were then estimated using Henderson's method. We will denote the estimator based on this model and **POST** by **EBLUP2** and the estimator based on **GREG** by **EBLUP2***. In the second model we assume the variance component σ_ν^2 to be zero. The estimator based on **POST** and this second model will be denoted by **EBLUP1**, and that based on **GREG** will be denoted by **EBLUP1***. Note that the estimators **EBLUP2** and **EBLUP2*** are optimal, in the sense that they are based on a correctly specified model, while **EBLUP1** and **EBLUP1*** are pseudo-optimal.

For the benchmarked estimators the benchmark was taken to be the estimator **EXP** at the SIC2 level. Ratio adjusted benchmarking was applied to all estimators. The two versions of MSE adjusted benchmarking were applied to the estimators **EBLUP2*** and **EBLUP1***, but not to any other estimators because of problems with estimated MSE matrices being singular. The MSE matrices of the **EBLUP** estimators were estimated by the "naive" estimator, *i.e.* $\boldsymbol{V} - \boldsymbol{V}(\boldsymbol{V}+\boldsymbol{W})^{-1}\boldsymbol{V}$ with $\boldsymbol{V}$ and $\boldsymbol{W}$ replaced by estimates.

3.2 Evaluation Measures

Suppose $\boldsymbol{m}$ simulations are performed in which $\boldsymbol{m}_1$ sets of different vectors of realized sample sizes for SIC3s by strata are replicated $\boldsymbol{m}_2$ times. The following measures can be used for comparing performance of different estimators. Let $\boldsymbol{i}$ vary from 1 to $\boldsymbol{m}_1$ and $\boldsymbol{j}$ from 1 to $\boldsymbol{m}_2$.

(i) Absolute Relative Bias.

$$\mathrm{ARB}_a = \left| m^{-1} \Sigma_i \Sigma_j (\mathrm{est}_{ija} - \mathrm{true}_a) / (\mathrm{true})_a \right| \tag{3.1}$$

The average of $\mathbf{ARB}_a$ over domains $\boldsymbol{a}$ will be denoted by **AARB**.

(ii) Root Mean Square Conditional Relative Bias.

$$\mathrm{RMSCRB}_a = \{ m_1^{-1} \Sigma_i (m_2^{-1} \Sigma_j \mathrm{est}_{ija} - \mathrm{true}_a)^2 / \mathrm{true}_a^2 - B \}^{1/2} \tag{3.2a}$$

$$B = m^{-1}(m_2 - 1)^{-1} \Sigma_i [\Sigma_j \mathrm{est}_{ija}^2 - (\Sigma_j \mathrm{est}_{ija})^2 / m_2] / \mathrm{true}_a^2 \tag{3.2b}$$

The correction term $\boldsymbol{B}$ adjusts for bias in the first term due to $\boldsymbol{m}_2$ being finite. **ARMSCRB** will denote the average of $\mathbf{RMSCRB}_a$ over areas $\boldsymbol{a}$.

(iii) Mean Absolute Relative Error.

$$\mathrm{MARE}_a = m^{-1} \Sigma_i \Sigma_j \,|\, \mathrm{est}_{ija} - \mathrm{true}_a \,| / \mathrm{true}_a \tag{3.3}$$

and **AMARE** denotes the average of $\mathbf{MARE}_a$ over domains $\boldsymbol{a}$.

(iv) Relative Root Mean Square Error.

$$\mathrm{RRMSE}_a = \{ m^{-1} \Sigma_i \Sigma_j (\mathrm{est}_{ija} - \mathrm{true}_a)^2 \}^{1/2} / \mathrm{true}_a \tag{3.4}$$

and **ARRMSE** as before denotes the average over domains.

The precision (i.e. the Monte Carlo standard error) of each measure depends on m_1, m_2. It can be seen that for all measures except (ii), the optimal choice of m_1, m_2 under the restriction that $m_2 > 1$ is $m_1 = m/2$, $m_2 = 2$, since this minimizes the Monte Carlo standard error. For the second measure, the appropriate choice of m_1, m_2 is less straightforward. For our simulation study we set $m_1 = 5000$, $m_2 = 2$.

3.3 Empirical Results

Figures 1 to 5 display the average evaluation measures from the Monte Carlo simulations for most of the estimators included in the study.

Figure 1 shows evaluation measures for unbenchmarked direct estimators. Clearly use of the covariate has a very beneficial effect in this example, as would be expected because of the model used to generate the data. The estimator **POST** is best among those which do not use the covariate, while **SRAT** and **GREG** are both best among those using the covariate.

Figure 2 shows the effect of combining the **POST** and **GREG** estimators with a regression synthetic estimator and compares the three methods of composite estimation. Generally, composite estimation shows some improvement in the evaluation measures **AMARE** and **ARRMSE** and some deterioration in the bias measures (**AARB** and **ARMSCRB**), with the **EBLUP**s showing a stronger effect than the **SSD**s. In this study there is very little difference between the two **EBLUP**s. The performance of the pseudo-optimal estimators, **EBLUP1** and **EBLUP1***, is the same as that of the optimal estimators, **EBLUP2** and **EBLUP2***, respectively; however, see also Figure 5 and the discussion below.

Comparing Figure 3 to Figure 2 we see the effect of benchmarking. Generally the effect of benchmarking here is a slight improvement in the overall bias (**AARB**) at the cost of some deterioration with respect to the other evaluation measures. The relatively poor performance of the benchmarked estimators is not surprising since the benchmark **EXP** performs relatively poorly; see Figure 4. Benchmarking would be expected to improve performance only in the case of serious model breakdown.

Figure 5 compares the three different methods of benchmarking. For the estimator **EBLUP1*** all three methods perform about the same. For **EBLUP2*** the ratio adjusted benchmarking method performs as well as for **EBLUP1***; however, the MSE adjusted methods perform more poorly. A possible explanation is that, with the extra variance component in the model underlying **EBLUP2***, the estimate of the MSE of **EBLUP2*** is of poor quality.

Acknowledgement

We are grateful to Dave Binder, Dave Dolson, and Jon Rao for helpful discussions. The second author's research was partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada held at Carleton University.

References

Battese, G.E., Harter, R.M., and Fuller, W.A. (1988). An error-components model for prediction of country crop areas using survey and satellite data. *Journal of the American Statistical Association*, **83**, 28-36.

Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistics Association*, **74**, 269-277.

Pfeffermann, D., and Barnard, C.H. (1991). Some new estimators for small area means with application to the assessment of farmland values. *Journal of Business and Economics Statistics*, **9**, 73-84.

Pfeffermann, D., and Burck, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, **16**, 217-237.

Platek, R., Rao, J.N.K., Särndal, C.E., and Singh, M.P. eds (1987). *Small Area Statistics: An International Symposium*; New York; John Wiley & Sons.

Rao, J.N.K. (1986). Synthetic estimators, SPREE and best model-based predictors of small area means. Technical Report, Laboratory for Statistics and Probability, Carleton University, Ottawa.

Rao, J.N.K., and Choudhry, G.H. (1993). Small area estimation: overview and empirical study. *Monograph proceedings of the International Conference on Establishment Surveys, Buffalo, June 1993*, to appear.

Särndal, C.E., and Hidiroglou, M.A. (1989). Small domain estimation: a conditional analysis. *Journal of the American Statistical Association*, **84**, 266-275.

Singh, A.C., and Mantel, H.J. (1991). State space composite estimation for small areas. Proceedings of *Symposium 91: Spatial Issues in Statistics*, Statistics Canada, Ottawa, November 1991, 17-25.

Symbols used in figures

△	**AARB**	– average absolute relative bias
○	**ARMSCRB**	– average root mean squared conditional relative bias
▽	**AMARE**	– average mean absolute relative error
□	**ARRMSE**	– average relative root mean squared error

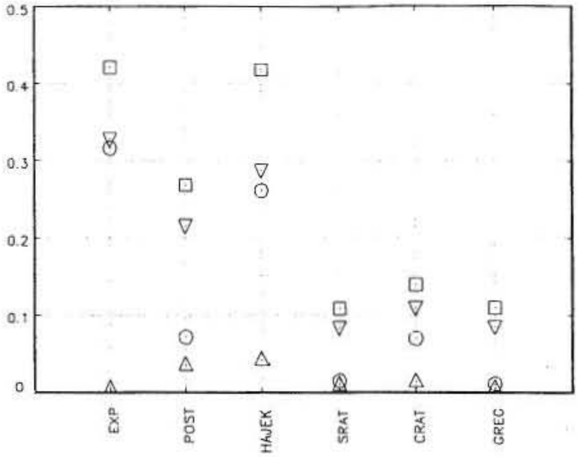


Figure 1: unbenchmarked direct estimators

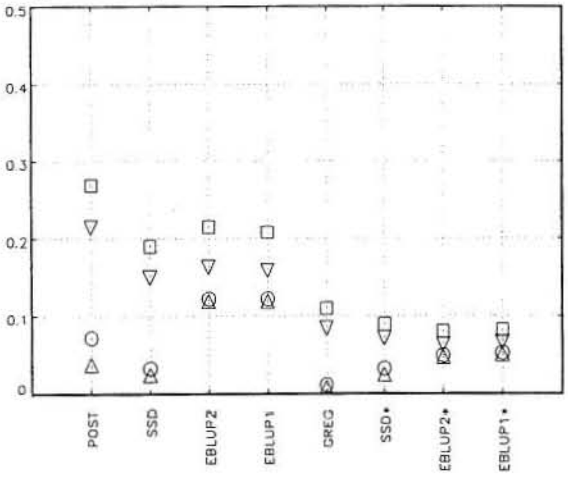


Figure 2: comparison of direct and composite estimators

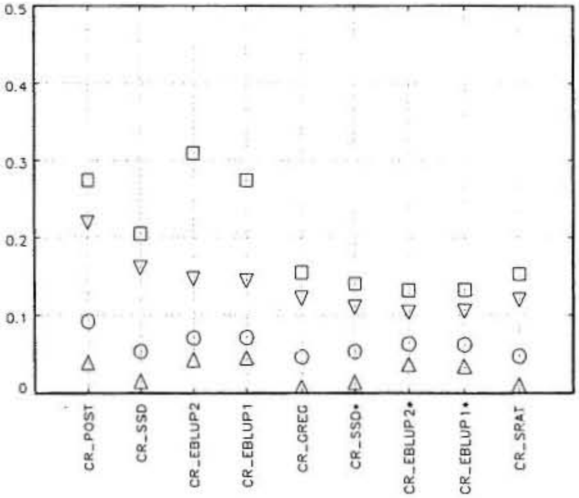


Figure 3: benchmarked direct and composite estimators

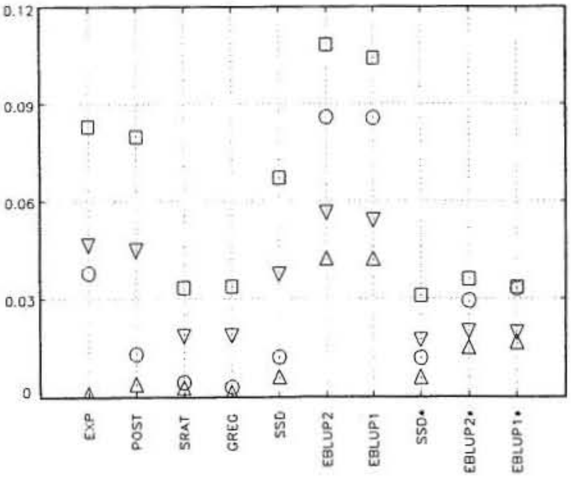


Figure 4: unbenchmarked estimators aggregated to large areas (SIC2 level)

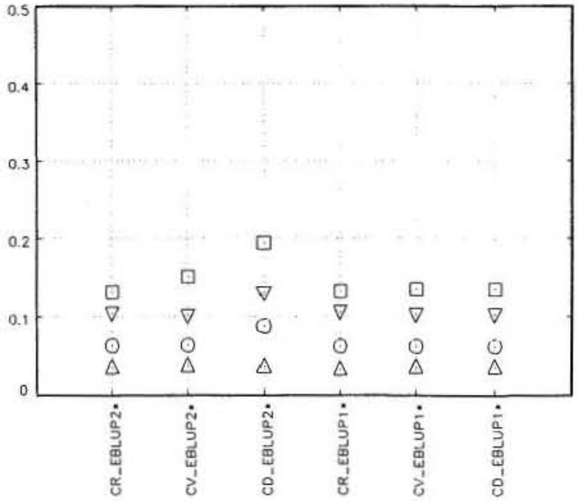


Figure 5: comparison of three benchmarking methods