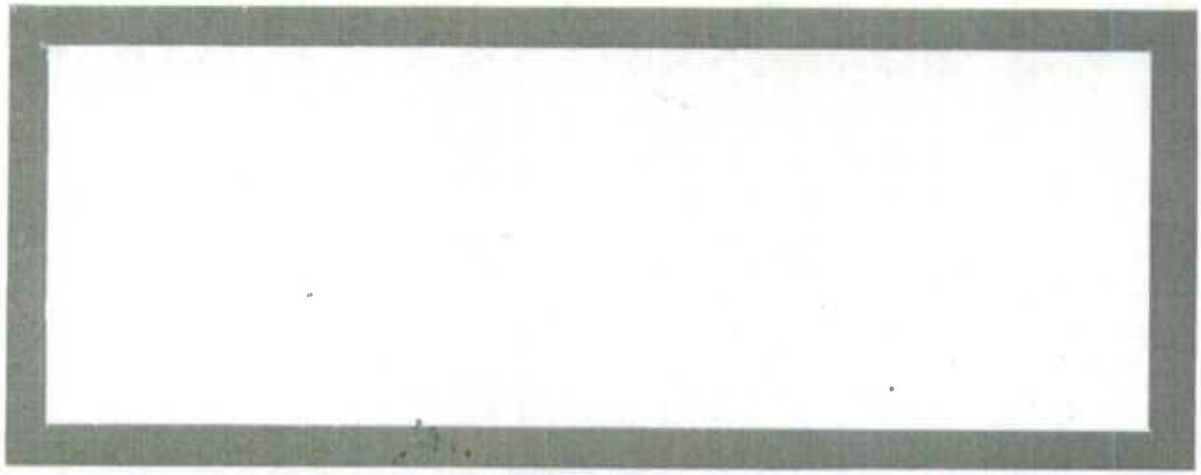
Statistics Canada Statistique Canada



Methodology Branch

Business Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes entreprises

Canada

WORKING PAPER NO. BSMD-94-003E

METHODOLOGY BRANCH

CAHIER DE TRAVAIL NO. BSMD-94-003E

DIRECTION DE LA MÉTHODOLOGIE

A SUMMARY OF THE ESTIMATION METHODOLOGY OF THE SURVEY OF EMPLOYMENT, PAYROLLS AND HOURS (January 1983 to February 1994)

by

D. Dolson

March 1994

A SUMMARY OF THE ESTIMATION METHODOLOGY OF THE SURVEY OF EMPLOYMENT, PAYROLLS AND HOURS (January 1983 to February 1994)

by

David Dolson
Business Survey Methods Division

December 1990
revised August 1992
revised March 1994

A SUMMARY OF THE ESTIMATION METHODOLOGY OF THE SURVEY OF EMPLOYMENT, PAYROLLS AND HOURS (January 1983 to February 1994)

David Dolson
Statistics Canada

ABSTRACT

At its inception in March 1983 the Survey of Employment, Payrolls and Hours (SEPH) used an almost unbiased expansion type estimator for estimation of totals. Because of a number of key events since then, the estimation methodology has undergone a series of changes. These changes are noted and non-technical descriptions of the revised estimation methods are given.

Starting with estimates for April 1992, estimates have been published using the 1980 Standard Industrial Classification (SIC) instead of the 1970 SIC which had been used since SEPH's beginning in March 1983. When the April 1992 estimates were published in June 1992, an historically revised series for the period January 1983 to March 1992 was published using the 1980 SIC. A non-technical description is given of the historical revision procedures.

RÉSUMÉ

À ses débuts en mars 1983, l'Enquête sur l'emploi, la rémunération et les heures (EERH) a utilisé un estimateur presque sans biais pour l'estimation des totaux. À cause de quelques évènements clés qui se sont produits depuis, la méthodologie d'estimation a subi une série de changements. Ces changements sont présentés et les nouvelles méthodes d'estimation sont décrites d'une façon non-technique.

Commençant avec les estimations d'avril 1992, les estimations sont publiées en utilisant la Classification type des industries (CTI) de 1980 au lieu de la version de 1970 qui a été utilisée jusque là. Lorsque les estimations pour avril 1992 etaient publiées en juin 1992, un ensemble d'estimations révisées pour la période historique a été publié en utilisant la CTI de 1980. Ce rapport donne une description non-technique des procédures de la révision historique.

A SUMMARY OF THE ESTIMATION METHODOLOGY OF THE SURVEY OF EMPLOYMENT, PAYROLLS AND HOURS (January 1983 to February 1994)

David Dolson, Statistics Canada
Business Survey Methods Division, 11th floor
R.H. Coats Bldg, Tunney's Pasture, Ottawa, Ontario, K1A 0T6
Statistics Canada

1 INTRODUCTION

At its inception for March 1983 reference month the Survey of Employment, Payrolls and Hours (SEPH) used an almost unbiased expansion type estimator for estimation of totals. Since then the estimation methodology underwent a number of changes because of a series of key events having an impact on the survey. Not all of these changes have been simple and easy to understand. Although each has been documented, there has been no single paper telling the entire story. Thus, getting the complete picture of what has transpired has not been easy. This paper serves to fill in that gap by documenting, in a non-technical manner, the various estimation methodologies used in SEPH over the period January 1983 to February 1994.

In March 1994 a major redesign of the survey was implemented. This included a substantial redesign of the frame and the estimation methodology that does not "qualify" as a modification to the methodology implemented in 1983. Hence it is not covered in this paper.

As of June 1992 with the release of estimates for April 1992 reference month SEPH started publishing its estimates using the 1980 Standard Industrial Classification (SIC) instead of the 1970 version. At that time SEPH's "historical revision" was released. The revised time series incorporated corrections for major data discontinuities which were present in the previously published estimates. This paper will also describe the procedures used to make these corrections, many of which were closely related to the same key events noted above.

The next section gives a brief overview of the SEPH design. Section three describes the estimators which were used in current production while section four describes the procedures used in the historical revision. For easy reference the appendix gives a list of the key events, terse descriptions of the estimators used in production in each time period, and terse descriptions of the correction factors applying to each period of time.

Finally, unless otherwise indicated, all dates in the remainder of this paper will refer to a reference month rather than a "real time" month. Also, the term "unit" will often be used as a synonym for establishment.

2 THE SEPH DESIGN

The SEPH is Canada's establishment based employment survey. It is designed to measure monthly levels and month-to-month trends of employment, paid hours and earnings at detailed industrial and geographic levels. All industries are covered except agriculture, fishing and trapping, private household services, religious organizations and military services. The design is described in detail by Schiopu-Kratina and Srinath (1986).

The sampling unit is the establishment. It is defined to be a production unit or grouping of production units which do not cross provincial boundaries and for which records provide data on the value of output, the cost of principal inputs and on the cost and quantity of labour. Currently there are about 800,000 in-scope establishments. This population is stratified by industry division (16), province (12) and estimated employment (4). Establishments are classified to employment size group according to their maximum estimated employment over a twelve month period. The size groups are labelled 1, 2, 3, and 4 and include establishments with maximum estimated employment of 1-19, 20-49, 50-199 and >199,respectively. Each stratum is further subdivided into cells defined by three digit SIC.

There are about 66,000 establishments in sample. Of these, about 34,000 are in sample with certainty (take-alls); this includes all establishments belonging to enterprises having 200 or more employees as well as a few other smaller categories of establishments. The take-some sample size is determined via sampling fractions which are defined for each stratum and which remain constant over time. The sample for each stratum is in turn allocated to cells proportional to the number of take-some establishments in each cell. Sampling fractions are constrained to be at least 1/100. The sample size in each cell is constrained to be at least min(3, the take-some population in the cell).

These sampling fractions were determined by specifying: a target coefficient of variation of three percent for the estimate of total employment, including that of take-all units, at the industry division by province level; and that the take-some sample be allocated to strata within each industry division by province proportional to the estimated total employment in the take-some population of the stratum. By applying this procedure prior to SEPH's start-up take-some sample sizes were determined. These in turn were converted to the sampling fractions which have been used each month as described above.

Partial rotation of the sample takes place every month. This is controlled at the cell level. Every sampled take-some establishment is in the sample for at least one year, with the exception of births. Upon rotating out of the sample, an establishment is kept out of the sample for at least one year. In-sample units are assigned to thirteen rotation groups, the number indicating the month in which most (the exceptions are births) units in the group rotated into the sample. Rotation group thirteen is for units which have been in sample for over twelve months and are eligible to rotate out of the sample but are prevented from doing so for lack of replacement units. When a unit rotates out of the sample it is placed in the "waiting" group for twelve months and is not eligible for reselection during that period. After this time, the unit moves to the "waiting for selection" group and is eligible for selection.

Births are sampled at the appropriate rate and are distributed at random to rotation groups one to twelve. Unselected births are assigned to both waiting and waiting for selection so that they are correctly represented in each.

Deaths are identified from both the sample and from external sources. These are removed from both the sample and the out-of-sample populations each month. An adjustment, called the death adjustment factor, is made to account for the exclusion of deaths from the sample.

Most data collection is by mail, with the majority of the remainder by telephone. Non-response and edit failures are followed up by telephone. Once the period for follow up is over, complete record imputation takes place for any remaining non-respondents and for questionnaires having uncorrected edit failures.

3 ESTIMATORS USED IN CURRENT PRODUCTION

3.1 March 1983 to September 1987

The estimator described by Schiopu-Kratina and Srinath (1986) was used. Estimation of totals and variances is done at the cell level and these are aggregated to the desired level. Outliers are given a weight of one and the weights for other establishments in the same cell are adjusted accordingly. Totals are estimated via a nearly unbiased expansion type estimator. The increase in the level of employment estimates (about 6%) due to the change of frame from the Labour Division Master File (LDMF) to the Statistics Canada Business Register (BR) in January 1987 was reflected in the published estimates.

3.2 October 1987 to March 1988

In October 1987 the size classifications of about 27,000 units on the BR were changed; this was about 3.8% of the units in the SEPH frame. Conceptually, new size codes were available for all units even though in most cases they were the same as the old size codes. On SEPH's frame file the size codes of out-of-sample units were updated immediately while the size codes for take-all units were not updated at all. Units which were take-somes and became take-alls (increase in size) were made take-alls immediately. The size codes for remaining in-sample take-some units were updated when the units rotated out of the sample. Sample selection and rotation continued on the basis of the *current* size codes. (This is the usual SEPH procedure for the much smaller number of size code changes that occur each month.) In this context the current size code was the old size code for take-all units, while it was the new size code for out-of-sample units and units which either rotated into or out of the sample from October 1987 onwards. For any take-some unit which was in-sample in September 1987 the current size code was the old size code until it rotated out of the sample. An effect of this is that in-sample units classified by their old size codes were gradually replaced over a one year period by rotating in units classified by their new size codes. This process would have taken place over a longer period of time for cells in which units had been in sample for longer than one year; these are cells in which rotation takes place more slowly than is usual for SEPH.

Use of the expansion type estimator continued, although adjusted weights were used as described by Laflamme (1988). Weights for take-all units continued to be one. The weights for take-some units were computed on the basis of the current size codes for in-sample units and old size codes for out-of-sample units. Thus in October 1987, weights for take-some units were based almost entirely upon the old size codes; only units which had just rotated in would be classified according to their new size codes. (Of these only about 4% would have had old and new size codes which differed.) This was a nearly unbiased procedure. However, over the following year, the current size codes of more and more of the in-sample units were their new size codes. So, after a year, take-some weights were based upon old size codes for out-of-sample units and many units in cells having units which had been in sample for over a year and new size codes for all other in-sample units - a biased procedure, the bias arising solely from the take-some portion of the sample. For example, considering units with current size code 2, if the mean of units for which this is the old size code is different from that of units for which this is the new size code then a bias would result.

This weighting procedure was originally seen as a short term solution. However, its use continued up to the September 1990 estimates. A principal reason for this is that a series of delays in SEPH's start-up with the new Central Frame Data Base (CFDB), also known as the new Business Register or currently simply as the Business Register, as its frame occurred such that this start-up took place in October 1990, much later than had originally been foreseen.

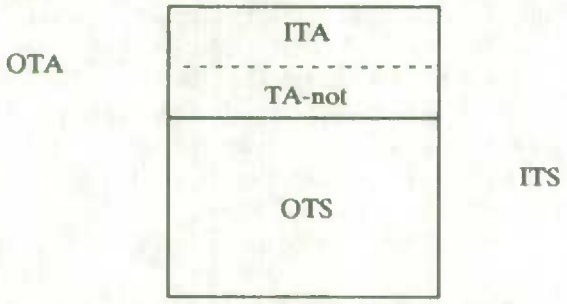
3.3 April 1988 to September 1990

Until March 1988 the take-all definition for SEPH included all units belonging to companies having more than one employment reporting unit (eru) - multi-establishment companies and single establishment companies with more than one eru. However, by early 1988 the SEPH sample size had increased substantially and it was necessary to reduce it. It was decided to restrict the take-all definition to do this.

The interim take-all definition revised the old definition by excluding existing multi-establishment companies with maximum employment less than 175. New multi-establishment companies with maximum expected employment less than 200 and additional establishments to multi-establishment companies previously excluded from take-all were also excluded on a monthly basis. All of these units, dubbed take-all-nots (TA-nots), were moved to the waiting group and were not eligible for selection until SEPH's change of frame from the old BR to the CFDB in October 1990. However, single establishment companies, of any size, with multiple erus remained take-alls because of system limitations. Some take-alls which would have been excluded on this definition were retained as take-all: units belonging to companies with less than 200 employees and with at least one establishment in manufacturing size group three, education units, units which would not have been adequately represented by in sample take-somes.

SEPH was publishing both an all units series, based upon the complete sample, and a take-all (old definition) series, based only upon the take-all units. With the implementation of the interim take-all definition not all units previously covered in the take-all series remained in the sample. It was decided to project forward the take-all series based upon the units included in the interim take-all. This methodology is described by Brown (Aug 1988) and Dupuy (April 1988). It was originally planned to use this procedure for only a short period of time until CFDB integration. However, due to delays occurring in starting use of the CFDB for SEPH this approach was used for much longer than expected.

In the remainder of this section I will describe in more detail what was done. I will use the following notation. First, the population can be split into two parts: old take-all units (OTA) and old take-some units (OTS) according to the old take-all definition. The OTA can be partitioned into two groups: the TA-nots, defined above, and take-all units according to the interim take-all definition (ITA). Finally, the interim take-some (ITS) is comprised of the TA-not and the OTS.



For April 1988, the estimate was split into three components: ITA, TA-not and OTS. Estimates for units in the ITA and the OTS were computed as before. (Weights for the OTS were computed excluding units in the TA-not group.) The TA-not estimate was computed as the simple average of two estimates. The first was computed by

$$TA\text{-}not_{1,apr} = TA\text{-}not_{mar}(ITA_{apr}/ITA_{mar})$$

where the subscript denotes the month. The second was computed by

$$TA\text{-}not_{2,apr} = \{ITS_{apr} - OTS_{apr}\} + \{(ITA_{mar} + ITS_{mar}) - (OTA_{mar} + OTS_{mar})\}.$$

Thus $TA\text{-}not_{apr} = (TA\text{-}not_{1,apr} + TA\text{-}not_{2,apr})/2$. Next the take-all estimates for April were computed as $ITA_{apr} + TA\text{-}not_{apr}$ and estimates for the all unit series were computed by $ITA_{apr} + TA\text{-}not_{apr} + OTS_{apr}$. These computations were done at three levels of aggregation: three digit SIC by province, one digit SIC by province and the industrial aggregate. Finally, adjustments were made to estimates at lower levels of aggregation to ensure additivity to higher levels with emphasis being placed on preserving trends at the higher levels.

In subsequent months it remained necessary to produce estimates for both the take-all series and the all-units series. For the take-all series, ie. the OTA group, the estimate for month m was computed by $OTA_{m-1}(ITA_m / ITA_{m-1})$. For the all-unit series, the estimates for month m were computed by $All\text{-}unit_{m-1}\{(ITA_m + ITS_m)/(ITA_{m-1} + ITS_{m-1})\}$. These computations were done at three levels of aggregation: three digit SIC by province, one digit SIC by province and the industrial aggregate. Finally, adjustments were made to estimates at lower levels of aggregation to ensure additivity to higher levels with emphasis being placed on preserving trends at the higher levels.

Since the all-unit series and the take-all series were being projected forward independently it was mathematically possible that a take-all estimate could become larger than the corresponding all-unit estimate. This happened only rarely. When it did, the all-unit estimate was set equal to the take-all estimate.

At this point it is notable that starting in January 1989 the SEPH frame, the old BR, was no longer being updated for business births. Resources required for this activity had been reassigned to work on the CFDB. A small number of reactivations, processed via an automated procedure, were taking place each month. Although these reactivated units are properly treated by SEPH in the same way as births, they are not really new business births. A reactivated unit is one which at one time was active, then became inactive for a period of time, and has become active again. Deathing of businesses continued on the basis of signals from Revenue Canada - Taxation only; survey feedback (via the CBS718) was no longer used. SEPH estimation continued as described above.

3.4 October 1990 to March 1992

Starting in October 1990 the sample and frame were based upon the CFDB, including the use of the 1980 SIC instead of the 1970. A new take-all definition was implemented; it was changed so that new multi-establishment enterprises with maximum expected employment less than 200 are take-somes instead of take-alls as they were before.

There were two objectives for the first CFDB based SEPH sample - to be representative of the new frame, and to retain as many units from the last old BR sample as possible. Mach (1991) describes in detail what was done. In the months preceding October 1990 several strategies were applied to facilitate maximizing this overlap of samples. For October, instead of the normal sample rotation, the Kish and Scott (1971) (K&S) procedure was applied. This procedure took old sampling fractions used with the old BR and by moving units into and out of the sample was to adjust it so that it properly represented the CFDB using new sampling fractions (derived by the procedure described in section 2). Two groups of units are of particular interest.

These are the two groups of small multi-establishment companies (ie. less than 200 employees) which resulted when the interim take-all definition was implemented in April 1988. The first of these is the TA-nots. They were moved from the waiting group to the waiting-for-selection group and were sampled. The second group is the single establishment multi eru companies with less than 200 employees. The systems constraints which kept these units within the take-all definition no longer applied with the CFDB as the frame for SEPH and, with the new take-all definition, they were moved to the take-some population.

The TA-nots were out of the sample with certainty. However, they were treated in the K&S as take-somes having non-zero old probabilities of selection, just like any other take-some; this procedure was biased. They were eligible for selection for the October sample with the correct new probabilities of selection. Since it seems reasonable to expect that there is no difference in the distribution by size of units belonging to multis as compared to units belonging to single establishment companies it also seems reasonable to expect that the magnitude of this bias was small.

The second group of interest is the single establishment multi eru companies with less than 200 employees. Since these units were take-alls in September 1990, and were to become take-somes under the new take-all definition they received a special treatment in the K&S processing. As a temporary measure, these multis were all retained in the October sample and were treated as take-alls for estimation purposes. These units were then removed from the sample in November and moved to the take-some waiting group and were not eligible for selection for 12 months. However, in computing the probabilities of selection to be used with the CFDB these units were properly counted in the take-some population. Then, in the K&S procedure itself, appropriate sample sizes of take-somes for use in the long term were achieved by using the probabilities of selection just noted and by moving into and out of sample *existing* take-some units. The effect of this is that in October the sample size was larger than necessary to achieve the target C.V.s because this group of multis has been retained in the sample. In November, when they were removed, the sample size was reduced to its normal level.

This procedure was biased since these units, although take-somes, had probability zero of selection until they subsequently moved out of the waiting group. However, the magnitude of this bias was likely small. A comparison of this group of multis with regular take-some units in the same cells showed there to be very little difference between the two groups with respect to employment, earnings and hours. A less biased treatment of the above multi-establishment units would have been to allow some of them to be selected to remain in the November sample as take-somes while placing the remainder in the waiting and waiting-for-selection groups. However, this would have been very awkward from a systems point of view and it was decided not to do so.

Although the survey design was based upon the 1980 SIC, estimates continued to be produced using the 1970 SIC and the take-all series continued to be according to the old (not interim, not new) definition. There were a number of reasons for this. One was that SEPH analysts required some time to analyze the behavior of CFDB based estimates vis a vis estimates based upon the old BR. Another was to help minimize the disruption for users of SEPH data. Finally, time was required to complete the work on the historical correction project.

The estimation methods used for this period of time were first described by Sampson (May 1989 and July 1989). Some aspects of the methods described in those notes were not implemented. The methods that were implemented are briefly and accurately described in the context of a description of job streams by Dupuy (Sept 1990).

3.4.1 October 1990

For the purpose of producing 1970 SIC estimates for October 1990, a reduced or matched sample (a.k.a. common units) approach which projected forward the published estimates from September was used for the take-some component of the population.

Reporting units which were in both the September and the October samples were identified by a matching process. For September, estimates were produced using its adjusted weights (described in section 3.2) further altered to reflect the reduced sample size.

By matching to previous months' files, the 1970 SIC classifications of matched units in the October sample were retrieved. Since the October sample was based upon the CFDB instead of the old BR like the September sample, the population counts from the CFDB for October were not used in this matched sample procedure. This excluded any effect due to much larger population counts on the CFDB. Instead population counts on the 1970 SIC basis for October were taken to be the same as those from September. This therefore also excluded any effects due to births and deaths. Thus the same weights were used for October as for September and expansion estimates were computed for October based upon the reduced sample and the 1970 SIC.

The October take-all estimate included not only the matched units but also some take-all units which did not match to the September sample. This group consisted of units which although not present on the BR were covered. They occur because of changed business profiling on the CFDB. It is also notable that some of the matching take-alls were not included in the take-all estimation for October. These were units which were take-somes in September but became take-alls for October; they were treated as take-somes for October estimation.

Finally, the estimates for October were computed by multiplying the ratio of these reduced sample estimates as described above by the published estimates from September. This was done for both the take-all and the all-units series.

$$Oct_{pub} = Sept_{pub}(Oct_{expansion} / Sept_{expansion})$$

These computations were done at four levels of aggregation: three digit SIC by province, two digit SIC by province, one digit SIC by province and the industrial aggregate. Finally, adjustments were made to estimates at lower levels of aggregation to ensure additivity to higher levels with emphasis being placed on preserving trends at the higher levels.

3.4.2 November 1990

In November, the take-all clean-up was done. Multi-eru companies which had entered the take-all group since implementation of the interim take-all definition in April 1988 and having less than 175 employees were removed from the take-all group and placed in the take-some group. They were put in the waiting group and remained out of sample for at least a year. For November 1990 the matched sample approach was again used, although the procedure differed somewhat from that used for October since both samples were based upon the CFDB.

Reduced samples were identified in the same manner as above. For both October and November population counts were taken to be those corresponding to October. This procedure therefore excluded any effects due to births and deaths. It was chosen because it would be consistent with estimates since January 1989 which did not account for births due to these births not being included

on the old BR. However, it was inconsistent in that estimates during this period of time did account for deaths while this procedure did not do so. This was judged to be acceptable since the effect due to births was considered to be much larger than that due to deaths. Weights were modified for the reduced sample size and aggregate estimates were computed for each month. These were then converted to 1970 SIC basis. October published estimates for the take-all series were projected forward to November according to change as measured by units in the new take-all (NTA) group, which excludes units belonging to multi-establishment companies with maximum expected employment less than 200.

$$OTA_{nov,\,pub} = OTA_{oct,\,pub}(NTA_{nov}/NTA_{oct})$$

Similarly the all-unit published estimates were projected forward on the basis of expansion estimates computed using the new take-all/take-some split.

$$All\text{-}unit_{nov,\,pub} = All\text{-}unit_{oct,\,pub}\{(NTA_{nov}+NTS_{nov})/(NTA_{oct}+NTS_{oct})\}$$

These computations were done at four levels of aggregation: three digit SIC by province, two digit SIC by province, one digit SIC by province and the industrial aggregate. Finally, adjustments were made to estimates at lower levels of aggregation to ensure additivity to higher levels with emphasis being placed on preserving trends at the higher levels.

Since the all-unit series and the take-all series were being projected forward independently it was mathematically possible that a take-all estimate could become larger than the corresponding all-unit estimate. If this happened the all-unit estimate was set equal to the take-all estimate.

3.4.3 December 1990 to March 1992

For production of 1970 SIC based estimates for this period, the approach was similar to that for November 1990 except that the full samples were used. Estimates for a reference month m as well as for the previous month, m-1, were computed on the 1980 SIC using the expansion estimator and the full sample in each month. Population counts were taken to be those from m-1 and thus the effect of births and deaths was excluded. These estimates were converted to 1970 SIC. The take-all series was projected forward.

$$OTA_{m,\,pub} = OTA_{m-1,\,pub}(NTA_{m}/NTA_{m-1})$$

Similarly, the all-unit series was projected forward.

$$All\text{-}unit_{m,\,pub} = All\text{-}unit_{m-1,\,pub}\{(NTA_{m}+NTS_{m})/(NTA_{m-1}+NTS_{m-1})\}$$

These computations were done at four levels of aggregation: three digit SIC by province, two digit SIC by province, one digit SIC by province and the industrial aggregate. Finally, adjustments were made to estimates at lower levels of aggregation to ensure additivity to higher levels with emphasis being placed on preserving trends at the higher levels.

Since the all-unit series and the take-all series were being projected forward independently it was mathematically possible that a take-all estimate could become larger than the corresponding all-unit estimate. If this happened the all-unit estimate was set equal to the take-all estimate.

3.5 April 1992 to February 1994

Labour Division started publication of estimates according to the 1980 SIC in June 1992 with the release of March final estimates and April preliminary estimates. Two series of estimates were produced - the all-units series and a large units series. The large units series replaced the take-all series and includes all enterprises having 200 or more employees. (All of these units are take-alls.) Two weeks earlier, in mid-June, the historically corrected series, described below and covering the period up to March 1992, was released.

The estimator is essentially the expansion type estimator described by Schiopu-Kratina and Srinath (1986). From April to June 1992 the estimator incorporated a post-stratification adjustment in a small percentage of SICs. The procedure is described in detail by Gossen (1992). This was to correct for a small bias in the sample representativity which was gradually being corrected via the sample rotation. As the sample was corrected, the post-stratification was used in fewer SICs and was eliminated entirely with release of final estimates for June 1992. More detail on this are provided in section 4.5.

4 HISTORICAL REVISION

When SEPH starts publication of estimates using the 1980 SIC it will release an historically revised series using the 1980 SIC. The objective of the historical revision project was to produce an historically consistent SEPH time series covering the period from March 1983 to March 1992. In addition estimates for January and February 1983 were added to the time series. (Although SEPH had been in production the estimates were not published at the time).

First, over the entire period corrections were made for anomalies which occurred from time to time in the data series. An example of such an anomaly would be an unusual change in an estimate due to the simultaneous rotation in of an unusually large unit and rotation out of an unusually small unit, within the same cell. This particular item is not discussed further in this paper; however the remaining items mentioned in the next few paragraphs are expanded upon later in this section.

For the period from March 1983 to September 1990 the starting point for the revision was the data which were published. The jump which occurred due to the LDMF-BR changeover was phased in over the period April 1983 to January 1987 when the frame change took place. A correction for the bias arising from use of the adjusted weights after the October 1987 size code updates was phased in over October 1987 to September 1988. As well, all estimates from October 1988 to September 1990 were adjusted upwards to include this change in level. Estimates for the period January 1989 to September 1990 were adjusted to account for the business births not included in the old BR, and hence not in the SEPH estimates, during that period. Last for this period, estimates for October 1987 to September 1990 were adjusted upwards accounting for some additional business births which would normally have appeared on the frame at that time but which did not. These units had remained as unclassified units until they were "discovered" late in the historical revision process and were birthed over February to August 1991.

For October 1990 to March 1992, the revised estimates are based upon expansion estimates and post-stratified estimates. For October 1990 to July 1991 an adjustment also had to be made to account for the old unclassified units which were birthed to the frame over February to August 1991.

Finally, the estimates prior to October 1990 were converted to the 1980 SIC and were adjusted upwards so that the transition across September/October 1990, ie. across the change in frame from the old BR to the CFDB, would reflect only economic change and not any other changes which would be due to the change in frame.

A large units series of estimates was derived. This was done by including all take-all units satisfying one of the following two conditions: (1) maximum employment over a calendar year exceeds 199, (2) maximum employment over a calendar year exceeds 174 and maximum in the previous year over 199. Take-some units which might have grown to that size were not included.

4.1 LDMF to BR Frame Change (January 1987)

The first adjustment was for the roughly 6% jump in estimated totals which occurred in January 1987 due to the frame change from LDMF to BR. This jump is measured by the ratio between BR based and LDMF based estimates for January. It was assumed that the LDMF and BR were consistent with each other in March 1983 when SEPH started. It was further assumed that this discrepancy arose at a uniform rate over some number of months up to January 1987.

Brown (March 1988) researched how big this number should be, considering intervals ranging in length from 6 to 45 months. Huot (1988) further considered the issue using time series methods. Finally, it was decided that the number of months should be 45; this would leave the March 1983 estimates and the January 1987 estimates unchanged and phase in the jump over the period April 1983 to December 1986.

The jump was then "wedged in" over those months. Let D_y be the relative increase in the estimated total of y measured in January 1987. Letting i = 1refer to April 1983 and i = 45refer to December 1986, in month i the corrected estimate for the total of y was

$$\hat{y}_{i,\,corrected} = \hat{y}_{i,\,published}(1 + D_y(i/45)) \qquad i = 1, \ldots, 45.$$

This procedure was applied at the detailed level only - three digit SIC by province. In general this provided good results. In cases where it did not analysts derived a revised series by hand, working from the original published estimates. Estimated totals at aggregated levels were computed by adding up estimates at detailed levels.

Ratios like average weekly earnings and average weekly hours were then recomputed using these corrected estimates.

4.2 BR Size Reclassification (October 1987)

The next item in the historical revision was the BR size reclassification which took place in October 1987. This is discussed in detail by Dolson (Jan. 1991). From October 1987 to March 1988 published estimates had been computed using the expansion type estimator with adjusted weights in which old size codes were used for out-of-sample units and current size codes for in-sample units. From April 1988 to September 1990 estimates were computed via the projection approach described in section 3.3 and using the adjusted weights. Call the expansion estimates used in this procedure the *adjusted* (adj) estimates. As noted in section 3.2 this estimator was nearly unbiased in October 1987 but gradually became more biased downwards until September 1988 after which the relative bias due to this effect remained roughly constant.

From October 1987 onwards estimation could have continued using unadjusted weights, in which case new size codes would have been used for out-of-sample units and current size codes for in-sample units. Call these estimates the *unadjusted* (unadj) estimates. This approach would have produced upwardly biased estimates in October 1987 with the bias gradually shrinking towards a very small level by September 1988.

Ideally the magnitude of the bias in the published estimates would have been measured using September 1988 reference month. However, needed files were no longer available. So, the bias was measured using the closest available reference month for which all necessary files were available - November 1988. At the aggregate level the unadjusted estimates of totals were found to be approximately one percent higher than the adjusted estimates of totals.

This difference was wedged in over October 1987 to September 1988 by a technique similar to that noted above in section 4.1. Thus with i = 1referring to October 1987, the difference for a variable y was measured in November 1988 by $D_y = \hat{y}_{14,unadj} / \hat{y}_{14,adj} - 1$. Then the corrected estimate was produced by

$$\hat{y}_{i,corrected} = \hat{y}_{i,published}(1 + D_y(i/12)) \qquad i = 1, \ldots, 12.$$

Then for subsequent months, October 1988 to December 1988, corrected estimates were produced by

$$\hat{y}_{i,corrected} = \hat{y}_{i,published}(1 + D_y) \qquad i = 13, 14, 15.$$

These computations were done at four levels of aggregation: three digit SIC by province, two digit SIC by province, one digit SIC by province and the industrial aggregate. Finally, adjustments were made to estimates at lower levels of aggregation to ensure additivity to higher levels with emphasis being placed on preserving trends at the higher levels.

4.3 No Births on the BR (January 1989 to September 1990)

Published estimates from January 1989 to September 1990 did not account for business births. This was because the frame, the old BR, was no longer being updated for births. The methods to account for these missing births are documented in a draft report by Brown (July 1990). Several steps were involved.

The first was to identify the set of missing births. This was done using two components - active PD accounts on the non-integrated portion of the CFDB as of December 1989 but missing from the old BR, and births to the CFDB since January 1990. These were unduplicated with respect to each other as well as with respect to units on the old BR including the small number of births which had been appearing on the old BR.

Next, a month of birth to the SEPH frame had to be estimated for each of these units. A new business is normally birthed to the SEPH frame within a few months after it makes its first remittance to Revenue Canada - Taxation. A probability model was determined for the relation between first remittance month and SEPH birth month on the basis of SEPH births in October and November 1988. Having obtained the first remittance date for each missing unit, a SEPH birth month was hence assigned using this lag model.

Classification information available in late 1990 for these missing births was used. For some units this information was not available. Thus for each month there was a residual pool of unclassified births. For this group, counts of births by SIC were imputed so that counts of these births plus counts of those that were classified corresponded to the distribution of the SEPH births in 1988 by season.

Thus counts of missing births were available each month by cell. It was decided to account for these births by adjusting population counts in the weights. This would effectively impute an average size by cell of SEPH respondents to the missing births. However, because births are smaller on average than older units and because none of these missing births had been sampled, the counts of births were reduced to account for this. This aspect of the model was estimated by age, so that the growth of birth units was reflected. PD remittance data were used to estimate this difference in size of births relative to older units. The following ratio was computed for each industry division at the Canada level for m = 1,..., 19

$$r = \frac{(average\ \ remittance\ \ of\ \ units\ \ less\ \ than\ \ m\ \ months\ \ old)}{(average\ \ remittance\ \ of\ \ units\ \ more\ \ than\ \ m-1\ \ months\ \ old)}$$

The values for r did not show a smooth growth towards one as one might expect; instead a number of slight irregularities were observed. Smoothed values for r were computed separately for each industry division using a simple regression model of the form $r = \hat{\alpha} + \beta m$.

Thus, each month starting in January 1989 a cumulated number of missing births, reduced according to the appropriate value of r, had been computed for each cell. Next, it was necessary to account for the deaths of these missing births. They were identified by examining the remittance data for these units. Then the cumulated counts of births were appropriately reduced to account for them.

Then the original weights for January 1989 to September 1990, unadjusted for the October 1987 size code changes, were modified by increasing the population counts by the above cumulated and reduced counts of missing births. So, using these modified weights, expansion type estimates $\hat{y}_{i,\,mod}$ could be computed for each month i. It was also necessary to do this for December 1988 although the weights were the original unadjusted unmodified weights. Then, with i = 1referring to January 1989, the corrected estimates were computed by

$$\hat{y}_{i,\,corrected} = \hat{y}_{i-1,\,corrected}(\hat{y}_{i,\,mod} / \hat{y}_{i-1,\,mod}) \qquad i = 1 \ldots, 21.$$

This was done at each of the four levels of aggregation (three digit SIC by province, two digit SIC by province, one digit SIC by province and the industrial aggregate) and adjustments made for additivity.

4.4 Old Unclassified Units (October 1987 to September 1990)

Over the period February 1991 to August 1991 a set of old unclassified units were birthed to the CFDB. With normal birth processing these units would have been present on the CFDB in October 1990. In addition, at about that time a group of births pertaining to 1988 which had been mistakenly excluded from the birth revision project (section 4.3) were discovered. The treatment of these units was described by Dolson (June 1991).

First, using the lag model described in section 4.3, a birth month for each unit was estimated on the basis of its month of first remittance. Then these units were partitioned into two sets: those with birth month in 1988 or earlier and those with birth month in 1989 or later.

Next, the first group was considered. Because they were births, their count was reduced by about 50% in the same manner as the counts of births in the birth revision project (section 4.3). Their impact was measured by increasing population counts in the original unadjusted weights for December 1988. Estimates using the original unadjusted weights, $\hat{y}_{Dec88,\,unadj}$, were already available. Estimates in which the original unadjusted weights were increased by including in the population counts the reduced counts

of these missing units, $\hat{y}_{Dec88.\,excl}$ were also computed. Then their impact was measured by $D_y = \hat{y}_{Dec88.\,excl} / \hat{y}_{Dec88.\,exadj} - 1$. Then this correction was implemented by further correcting the estimates resulting from the procedures described in sections 4.2 and 4.3. So corrected estimates for October 1987 to December 1988 were produced by wedging in this difference over the fifteen month period, i = 1referring to October 1987:

$$\hat{y}_{i.\,corrected2} = \hat{y}_{i.\,corrected}(1 + D_y(i/15)) \qquad i = 1, \ldots, 15$$

Then for subsequent months, January 1989 to September 1990 corrected estimates were produced by

$$\hat{y}_{i.\,corrected2} = \hat{y}_{i.\,corrected}(1 + D_y) \qquad i = 16, \ldots, 36.$$

These computations were done at four levels of aggregation: three digit SIC by province, two digit SIC by province, one digit SIC by province and the industrial aggregate. Finally, adjustments were made to estimates at lower levels of aggregation to ensure additivity to higher levels with emphasis being placed on preserving trends at the higher levels.

Next, deaths in this set of births as well as the births for the latter period, 1989 or later, needed to be taken into account. It turned out that the number of such deaths was approximately equal to the number of these latter births. As well, their distributions by industry were very similar. It was decided that in the context of the approximations being made as well as because of the need for simplicity given the time constraints at the time, it would be reasonable to do nothing about either group thus allowing the effects of each group to roughly balance each other off.

4.5 Biased Sample (October 1990 to March 1992)

Starting in October 1990 the SEPH frame and sample were based upon the new CFDB. The size codes of many of the units in common with the old frame were different from what they had been on the old frame.

After a number of months, it was discovered that there was a bias in the SEPH sample that had been present since October 1990. Units whose size code had decreased were over-represented and units whose size code had increased were under-represented. Using SEPH's expansion type estimator, the effect was an over-estimation of totals. Via SEPH's usual sample rotation this sample bias would gradually self correct so that in succeeding months the over-estimation would shrink towards zero. In most cells the correction would take place over one year, while in cells where rotation only takes place very slowly because of the small size of the cell the correction would take somewhat longer.

In the published estimates using the projection approach described in section 3.4.3 the effect was one of exaggerated decline in estimates of totals. This was because, each month in addition to any economic effects, the change from the previous month included a reduction in estimated totals due to the bias being reduced via sample rotation.

For the historical correction for October 1990 to March 1992 and as a temporary measure for the published estimates for April 1992 to June 1992 a procedure of post-stratification was developed to correct for the sample bias. The methodology is described in detail by Gossen (1992); I will provide a very brief summary here.

Take-some units are post-stratified based on their September 1990 size code on the old BR and their October 1990 size code on the CFDB. The post-strata were defined as illustrated by the following diagram. For example, units whose size codes were 2 or 3 or 4 in September and 1 in October have post-stratum code 2. Births since October 1990 are treated as non-matches to the old BR. Post-stratification is not necessary for take-all units - primarily units in size group 4.

Sept. Size	October Size			
	1	2	3	4
1	1	3	6	XXX
2	2	4		XXX
3		5	7	XXX
4				XXX
Non-match	9	10	11	XXX

In principle it would be desirable for the post-stratification to take place at as low a level as possible for which the sample size is large enough to assure reliability. Post-stratification was done at the level of Canada by two digit SIC group. The grouping of two digit SICs was necessary primarily to ensure adequate sample size in the post-strata. The grouping was allowed to differ by October size group. Province was not accounted for in the post-stratification for the same reason - to ensure adequate sample size in the post-strata. There are a number of special cases and situations which had to be considered; these are described by Gossen (1992). Suffice to say that in each month for each take-some unit a post-stratification adjustment to its weight was determined as a function of its post-stratum code and the two digit SIC group to which it belonged.

The post-stratification was a temporary procedure for SEPH. Thus criteria were put in place to determine when to stop post-stratifying. For this purpose each two digit SIC group by size combination was considered independently. January 1992 was the first month in which the post-stratification was performed for only a subset of the cases in which it was originally applied. It was turned off in 65% of the cases in January, with this figure increasing to 74% by March 1992 and to 100% by July 1992.

Because the post-stratification was done at the Canada by two digit SIC group level an issue was whether estimates at lower levels should be considered to be sufficiently reliable. Initially many of the post-stratification adjustments were relatively large in magnitude. In general, these declined with time. By an evaluation of the post-stratified estimates vis a vis the simple expansion type estimates it was decided that for October 1990 to August 1991 the post-stratified estimates should be considered reliable down to the province by SIC2 group level and afterwards at all levels of aggregation.

For the earlier period it was therefore necessary to develop a methodology to produce estimates down to the province by SIC3 level. For all of these eleven months, each post-stratified estimate at the two digit SIC group by province level, $\hat{y}_j$, was disaggregated to its three digit SIC components, $\hat{y}_{k\epsilon j}$, according to the distribution shown in the common units projections of the historically revised series, $\bar{y}_j$ and $\bar{y}_{k\epsilon j}$.

$$\hat{y}_{k\epsilon j} = \hat{y}_j(\bar{y}_{k\epsilon j}/\bar{y}_j) \qquad \forall k \in j, \forall j$$

The process is described in some detail by Dolson (Nov. 1991).

4.6 Old Unclassified Units (October 1990 to August 1991)

Over the period February 1991 to August 1991 a set of old unclassified units were birthed to the CFDB. With normal birth processing these units should have been present on the CFDB in October 1990. This is the same set of units discussed in section 4.4.

For October 1990 to July 1991 these units were accounted for by modifying the original weights by increasing the population counts by the number of these births. Units whose first remittance date was January 1990 or later were added into post-strata 9, 10, or 11 as appropriate thus treating them more like births. (By so doing, it was not necessary to reduce their count as it was for the earlier period discussed in section 4.4.) Units with earlier first remittance dates were randomly assigned to the other post-strata within the size group based on the proportion of units already in each post-stratum. Over the latter months in this period these counts were reduced as the units actually appeared on the frame.

4.7 BR to CFDB Frame Change (October 1990)

The procedures described in sections 4.1 to 4.6 produced two series, one for March 1983 to September 1990 on the 1970 SIC basis and the other for October 1990 onwards on the 1980 SIC basis. Two things needed to be done. First the series for the earlier period were converted to the 1980 SIC basis.

The second was to adjust for any level difference between the two series. First, the historically adjusted estimates for September 1990 were projected forward to October 1990 using the published trends. These were then converted to the 1980 SIC. This was done at four levels of aggregation (three digit SIC by province, two digit SIC by province, one digit SIC by province and the industrial aggregate) and adjustments were made for additivity. Next, the ratios of the post-stratification based estimates for October to these projected estimates were computed. Then each estimate in the period March 1983 to September 1990 was multiplied by the appropriate ratio and adjustments made for additivity.

4.8 The Magnitude of the Revisions

The corrections of anomalies, although important at the detailed level, were very small when considered at the aggregate level.

The correction for the LDMF to BR frame change resulted in an upwards adjustment of published aggregate employment estimates for December 1986 of about six percent. Estimates for March 1983 were unchanged while estimates for the intervening months were adjusted upwards by gradually increasing amounts up to the six percent.

The correction for the use of the adjusted weights yielded increases in published employment estimates of zero percent in October 1987 gradually increasing to about one percent in September 1988. Estimates for October to December 1988 were also increased by this same percentage. This correction factor also resulted in an upwards adjustment of employment estimates for January 1989 to September 1990 by the same ratio.

Next was the correction for no births on the BR during January 1989 to September 1990. This resulted in a further adjustment upwards ranging from something just over zero percent in January 1989 to about 4.5 percent in September 1990.

The addition of the old unclassified units produced a further increase relative to published estimates of aggregate employment ranging from just over zero percent in October 1987 to about 0.6 percent in December 1988. This latter increase also applied to all months up to September 1990.

The above set of adjustments produced a consistent series over January 1983 to September 1990. The entire series was adjusted upwards by just less than one percent in order to smoothly match the higher estimates coming from the improved coverage of the CFDB.

In October 1990 the post-stratified estimate of aggregate employment was about three percent lower than the expansion estimate. (This would include the effect due to the addition of the old unclassified units.) By October 1991 it had become about 0.7 percent higher and by June 1992 the difference had converged back towards zero with the post-stratified estimate about 0.2 higher than the expansion estimate.

APPENDIX

KEY EVENTS	
March 1983	SEPH starts
January 1987	LDMF to BR frame change
October 1987	BR size reclassification
April 1988	interim take-all definition
January 1989	no births on the BR
October 1990	first CFDB based sample, new take-all definition
November 1990	take-all clean-up
February 1991	start birthing of old unclassified units
April 1992	start publishing by 1980 SIC

ESTIMATORS IN CURRENT PRODUCTION	
PERIOD	ESTIMATOR
March 1983 to September 1987	Expansion type
October 1987 to March 1988	Expansion type with adjusted weights
April 1988 to September 1990	Projection using adjusted weights and full sample
October 1990	Common units projection using weights from September 1990 altered to reflect the reduced sample
November 1990	Common units projection using weights from October 1990 altered to reflect the reduced sample, conversion to 1970 SIC
December 1990 to March 1992	Full sample projection using weights in which Ns for month m are taken from m-1, conversion to 1970 SIC
April 1992 to June 1992	Expansion type in most cases, post-stratified in some
July 1992 onwards	Expansion type

HISTORICAL REVISION CORRECTION FACTORS	
FACTOR	PERIOD AFFECTED
0. Append January and February 1983	January and February 1983
1. Anomalies occurring from time to time in the published series	March 1983 to March 1992
2. Wedging in of LDMF to BR frame change jump	April 1983 to December 1986
3. Adjusting upwards to account for bias due to use of adjusted weights (BR size reclassification)	October 1987 to December 1988 explicitly January 1989 to September 1990 implicitly
4. No births on the BR	January 1989 to September 1990
5. Old unclassified units	October 1987 to July 1991
6. Post-stratification based estimates	October 1990 to March 1992
7. Conversion to 1980 SIC	January 1983 to September 1990
8. Level adjustment to make transition from BR to CFDB based frames "smooth"	January 1983 to September 1990

HISTORICAL REVISION CORRECTION FACTORS	
PERIOD AFFECTED	FACTORS
January and February 1983	0, 7, 8
March 1983	1, 7, 8
April 1983 to December 1986	1, 2, 7, 8
January 1987 to September 1987	1, 7, 8
October 1987 to December 1988	1, 3,5, 7, 8
January 1989 to September 1990	1, (3), 4, 5, 7, 8
October 1990 to July 1991	1, 5, 6
August 1991 to March 1992	1, 6

REFERENCES

BROWN, A. (March 10, 1988). Choice of Transition Interval. *memo to R. Dupuy.*

BROWN, A. (August 2, 1988). Methodology to disseminate Estimates excluding impact of New Take-all Definition. *memo to R. Dupuy.*

BROWN, A. (July 1990). Methodology to revise SEPH Estimates for Births. *BSMD.*

DOLSON, D. (January 22, 1991). The Historical Correction Project and SEPH Estimates Following the Size Code Updates in October 1987: Some Notes from our Meeting on January 16, 1991. *memo to distribution.*

DOLSON, D. (June 27, 1991). Missing Births in 1988 etc., etc.: My Understanding of What we Decided to do. *memo to distribution.*

DOLSON, D. (Nov. 1, 1991). SEPH Estimates: 1983 to Present . . . Take 2. *memo to distribution.*

DUPUY, R. (about April 1988). Making the Take-all Redefinition Transparent to Users. *Labour Division.*

DUPUY, R. (Sept. 10, 1990). DIP Part 2 Jobstreams. *memo to distribution.*

GOSSEN, M. (May 1992). Post-stratification of SEPH Estimates. *BSMD.*

HUOT, G. (May 1988). Selecting Linkage Transition Intervals to Minimize Trend-cycle and Seasonal Distortions. *Methodology Branch working paper TSRA-88-013E.*

LAFLAMME, G. (April 28, 1988). Revised weight component files for cycle 710 to XXX. *memo to A. Brown.*

KISH, L. and SCOTT, A. (1971). Retaining Units After Changing Strata and Probabilities. *Journal of American Statistical Association*, 66, 461-470.

MACH, L. (July 1991). SEPH's Transition to the new Business Register: Application of 'Kish & Scott' System and the Resulting Sample. *BSMD.*

SAMPSON, G. (May 30, 1989). Projection of 1970 SIC Estimates: Weight Component Files. *memo to A. Brown.*

SAMPSON, G. (July 5, 1989). Requirements for LDSS to Produce Weight Component Files for Projection. *memo to A. Brown.*

SCHIOPU-KRATINA, I. and SRINATH, K.P. (July 1986). The Methodology of the Survey of Employment, Payroll and Hours. *BSMD.*