EXTENDING HISTORICAL COMPARABILITY IN INDUSTRIAL CLASSIFICATION

John S. Crysdale, Statistics Canada
11-B7 J.Talon Bldg., Ottawa, Ontario K1A 0T6

KEY WORDS: Automated industry classification, Standard Industrial Classification (SIC)

Weep not that the world changes--did it keep
A stable, changeless state, 'twere cause indeed to weep.
William Cullen Bryant (1824)

INTRODUCTION

The need to deal with changes in the basis of industrial classification is a perennial problem facing users of establishment-based data.

In Canada, the most recent such change involved the 1983 adoption of the 1980 version of the Standard Industrial Classification (SIC). As a result, the 171 manufacturing industries of the 1970 SIC, plus one non-manufacturing industry, converted to 236 1980 SIC manufacturing industries and three non-manufacturing industries. In many cases, the transition was simple: 79 1970 SIC industries converted on a one-to-one basis and two converted on a many-to-one basis. But, often, the transition was less simple: eleven 1970 SIC industries converted on a one-to-many basis and eighty converted on a many-to-many basis; in one case, a single many-to-many group comprised 59 1970 SIC industries and 84 1980 SIC industries.

The objective of this paper is to compare different ways--all fully automated--that a researcher with access to machine-readable microdata can deal with that classification break and put the data on a comparable basis. The paper deals with manufacturing establishments reporting detailed commodity data; in 1982 these accounted for 57.0% of statistical units and 96.0% of manufacturing activity shipments.

There are three basic strategies used here to achieve comparable classification: (1) Extending the 1970 SIC forward in time by applying it to establishments now classified on a 1980 SIC basis. This would enable researchers to update statistical work already undertaken on a 1970 SIC basis. (2) Extending the 1980 SIC backward in time by applying it to establishments now classified on a 1970 SIC basis. This would reflect the more current model of industry structure. (3) Finding equivalent aggregations of entire 1970 and 1980 SIC industries. The resulting industries are of neither the old standard nor the new, but are closely related to each.

Three methods are employed to extend the 1970 SIC and the 1980 SIC. These involve (1) using reported product detail, along with a set of resistance rules intended to prevent establishments from flip-flopping between industries, (2) using a forced one-to-one concordance and (3) using a mix of the two.

More than one method of reclassification exists, even with full access to the microdata, due to the subjective aspects of industry classification discussed in the next section. The one-to-one concordance implicitly reflects the subjective considerations embedded in the series from which reclassification is taking place. The product detail methodology must model them explicitly.

The first section of this paper deals with the classification process followed in creating the official series. The second section discusses, in general terms, three methods of extending SIC-based classification. In the third section, those methods are evaluated by using them to reclassify manufacturing establishments reporting commodity detail in 1982 and by then comparing the results against the official assignments for that year. In 1982, data were collected on a 1970 SIC basis, but published on both bases. The main finding is that, at the 4-digit level, the industry assignments which most closely match those of the official series are achieved by bringing the 1970 SIC forward and by doing so using a mix of methods. In the fourth section, a non-SIC strategy, aggregation, is discussed. That strategy is simple to apply; its main disadvantage is that the resultant industries are not as widely recognized as those of the SIC.

I. CLASSIFICATION IN THE OFFICIAL SERIES

Since much of this paper deals with replicating official industry assignments for manufacturing establishments reporting commodity detail, it is useful to review how those assignments are made.

Classification occurs at the 4-digit level of the SIC. Each 4-digit SIC industry is defined in terms of the manufacture of specific commodities which are said to be *primary* to that industry. At the establishment level, a tentative industry assignment is calculated by grouping reported commodity outputs by primary industry and by then determining which group accounts for the largest share of commodity shipments.

From the 1982 reference year to the present time, this calculation has been performed by machine. The result is then compared against the establishment's existing assignment (typically last year's code; or, for births, an assignment based on *nature of business* enquiries). If the comparison indicates that the subject establishment should be considered for transfer to another industry, a print-out is produced for manual inspection. This sometimes leads to an amendment to commodity codes or shipment values. If the existing and calculated industry assignments continue to differ, a number of subjective considerations enter the process to determine whether a transfer will be immediately implemented.

One such subjective consideration involves resistance rules. Such rules are intended to prevent establishments from being transferred as a result of small shifts in output proportions, unless those shifts are seen to be permanent. The effect on industry aggregates of transfers based on small changes is disproportionate. For example, if an establishment with shipments of $100 changes industry as a result of a $1 shift in output, the sending industry will decline by 100 times that $1 shift; and the receiving industry will increase by the same factor. If the shift is only temporary, and the transfer is reversed, the impact will be felt a second time. Detailed subject matter knowledge of industry conditions and intentions will limit such transfers. There is, however, no explicit set of rules.

Another subjective consideration involves industry coverage. On occasion, an establishment may be assigned to an industry that does not account for the largest share of that establishment's output. This can happen if the establishment is such a significant part of a given industry, that its exclusion would result in serious undercoverage of the industry's activities. Such treatment is more likely to occur if the industry accounting for the largest share of the subject establishment's output is one set up to incorporate otherwise unspecified activities, and if the subject establishment cannot be artificially split between the industries involved.

Classification may also be affected by confidentiality considerations. For example, if transferring a large establishment to a small, stable industry would effectively release its confidential data, the transfer might be postponed in order to permit publication of the data for that industry.

Size significance can also be a subjective consideration. A transfer may be postponed if an establishment is judged to have an insignificant impact on industry aggregates, especially if timeliness is at risk.

In summary, the official classification of manufacturing establishments reporting commodity detail is based on a mix of objective rules and subjective considerations.

II. EXTENDING THE SIC: GENERAL DISCUSSION

In this section, the three methods of extending SIC-based classification are discussed in general terms.

Method #1: Product Detail Coding

This method involves going to the microdata and calculating an industry assignment *from scratch*. It follows closely the process used to generate the official series. There are two differences.

The first difference involves the treatment of commodities reported at a level too aggregated to be to said to be primary to just one 4-digit industry. For example, services performed on goods owned by other manufacturers (custom and repair work) are covered by insufficiently detailed classes. In 1982, too-aggregated commodities accounted for just over 4% of the manufacturing activity shipments of establishments reporting commodity detail. In the official classification process, such activity is either made primary to the industry in which the reporting establishment is found or is made primary to *no industry*. That treatment requires that an industry assignment already exists or that manual intervention can occur. In the fully-automated approach used here, these commodities are either made primary to the target classification industry to which the reporting establishment is assigned by one-to-one coding, or are made primary to the target classification industry to which the reporting establishment was assigned in the previous year.

The second difference involves the subjective factors discussed in the previous section. Only resistance rules are explicitly modelled here. These have been codified so that the classification process can be fully automated.

In general terms the rules used are as follows: (1) If an establishment has experienced *significant* change, it is transferred immediately. (2) Otherwise, the transfer will be made when the change is seen to be *permanent*.

As applied here, change is measured as 100 minus the following:

$$\frac{\text{Value of current year shipments primary to the industry assigned in the previous year}}{\text{Value of current year shipments primary to the industry accounting for the largest share of current year activity}} \times 100$$

This formula produces values which range from 0 to 100. The greater the value, the greater the change. Change is considered *significant* if the value produced by the formula is greater than or equal to 67. The same threshold applies for reclassification to either the 1970 SIC or the 1980 SIC. And the change (however insignificant) is considered *permanent* if the calculated industry assignment of the subject establishment remains the same for two consecutive years; in such cases, transfer occurs in that second year.

Table 1: Incidence of Resistance Rules, 1982 Reclassification to the 1970 SIC

	Manufacturing Activity Shipments of Establishments with Commodity Detail %
Dominant SIC unchanged	93.5
Dominant SIC changed, test	
Delay transfer (< 67)	0.6
Transfer now (>=67)	0.5
Change persists, transfer	2.5
New, move to dominant SIC	3.0
Total	100.0

In order to link generated assignments to the official series of other years, reclassification to the 1980 SIC is performed *backward* through time; for the same reason, reclassification to the 1970 SIC is performed *forward* through time. This means that in implementing these resistance rules (and in handling too-aggregated commodities), *previous year* must be interpreted as the previous year in the reclassification process; it is not necessarily the previous calendar year.

To demonstrate the impact of this set of resistance rules, the error rates calculated later in this paper will be shown both before and after the rules are implemented. The *before* assignment is the same as is calculated by the automated edit, except for the differing treatment of too-aggregated commodities.

Method #2: Forced One-to-One Coding

This method involves reclassification from existing assignments by means of industry-level tables (see full version of this paper, reference [1]) that map each 1970 SIC to just one 1980 SIC, and each 1980 SIC to just one 1970 SIC. By way of example, 1970 SIC 2710 Pulp and Paper Mills, which splits into 1980 SIC 2711 Pulp Industry, 2712 Newsprint Industry, 2713 Paperboard Industry, 2714 Building Board Industry and 2719 Other Paper Industries, will be forced entirely to 1980 SIC 2712 (which accounts for the largest share of the value added of SIC 2710 in the cross-classified data of 1982). All establishments assigned to 1970 SIC 2710 will be recoded to 1980 SIC 2712, none will be recoded to 1980 SIC 2711, 2713, 2714 or 2719.

Forced one-to-one coding is perhaps the simplest way of effecting 4-digit reclassification. Access to and processing of detailed product data are not required. Any researcher with a list of an industry's constituent establishments can reclassify all those establishments. In fact, reclassification need not occur at the establishment level but can occur using published aggregates. Reclassification by this method also has the merit of reflecting subjective decisions embedded in the official series. For example, it reflects the application of resistance rules--without necessitating an explicit formulation of those rules. One limitation, is that, strictly speaking, a data-based concordance applies only to the year from which it was generated (although it would not typically be used to reclassify the data of that same year). And, even in that year, its application can produce errors of inclusion and exclusion (as can the other two methods of reclassification).

Method #3: Mix of Methods

This is a mix of forced one-to-one coding and of product detail coding (with resistance rules). It takes advantage of the one-to-one mapping in reflecting subjective considerations and of the product detail approach in mirroring actual practice.

Whether product detail or one-to-one coding is used for a given originating classification industry depends on whether that industry maps well (i.e., can be forced with less than some predetermined level of error, calculated as a percentage of its own shipments total, to a single class of the target classification). If so, forcing is used. Otherwise, the product detail approach is used.

As applied here, the error threshold is 3%. That level was selected after some experimentation. In mixed methods reclassification to the 1970 SIC, one-to-one coding handled 92.3% of subject shipments; for reclassification to the 1980 SIC, one-to-one coding handled 52.5% of subject shipments.

III. EXTENDING THE SIC: EMPIRICAL EVALUATION

In order to evaluate these methods, each is used to classify all establishments reporting commodity detail in 1982. Assignments are generated on both a 1970 SIC and a 1980 SIC basis. Those assignments are then compared against the official assignments of 1982, which also exist on both a 1970 SIC and a 1980 SIC basis. The official assignments are treated as correct. The method which most closely replicates the official 1982 series will be deemed best. It can then be used to extend SIC classification in other years.

Error Rate Measure

The error rate measure used here will be referred to as the *percent erroneously classified*. It ranges in value from zero to one hundred, and is calculated as:

$$\frac{\text{Erroneous inclusion + Erroneous exclusion}}{\text{Official-series shipments total + Methodology-based shipments total}} \times 100$$

Erroneous inclusion is the value of shipments of establishments wrongly included in a given industry by the subject methodology; erroneous exclusion is the value wrongly excluded from that same industry.

To illustrate the calculation of this measure, consider the hypothetical case where establishments officially classified to an industry report shipments of $100 and where the subject methodology assigns establishments reporting $110 to that same industry. Also, suppose that the shipments of establishments erroneously included in this industry total $40, and that those of establishments erroneously excluded total $30. Under these circumstances, the percent erroneously classified is 33.3--i.e., (($40+$30)/($100+$110))x100.

An alternative error measure involves comparing the shipments total of the official-series industry against that of the industry generated by the subject methodology, in this case, $100 and $110, respectively. This would indicate a 10% error rate. Such a comparison of aggregates neglects the establishment content behind those totals. Consequently, it can produce misleading results. For example, if instead of generating a shipments total of $110, the subject methodology had generated a total of $100, along with $100 of erroneous inclusion and $100 of erroneous exclusion, the alternative would have indicated zero error. The alternative is not used further.

Because data users often work at the 3- and 2-digit levels of detail, the various methodologies are also assessed at those levels, using the percent erroneously classified. This involves comparing the first three (or two) digits of the 4-digit code generated by the subject methodology against the corresponding digits of the official 4-digit code.

Results

Table 2 shows the percent erroneously classified evaluated at the 4-, 3- and 2-digit levels, averaged on a shipments-weighted basis to the all-manufacturing level (see reference [1] for error rates averaged at the 2-digit level).

The main conclusion arising from an examination of these data is that the best results are obtained by using mixed methods. Evaluated at the 4- and 3-digit levels for reclassification to either the 1970 SIC or the 1980 SIC, the mix outperforms the other methods.

Adding a set of resistance rules to the product detail methodology lowers error rates.

When evaluation occurs at higher levels of aggregation, the performance of all these methods improves. This is especially so for one-to-one coding, which improves very sharply between the 4- and 2-digit levels--indicating that most one-to-one error is internal to 3- and 2-digit industries. At the 2-digit level, one-to-one coding outperforms the mix of methods.

For reclassification to the 1980 SIC, one-to-one coding performs particularly poorly at the 4-digit level. Underlying the high error rate are 82 empty SIC classes (compared to 15 under the 1970 SIC) as well as all the erroneous inclusion to which such 100% erroneous exclusion corresponds. Those empty target classification industries exist as a result of imposing a one-to-one mapping on originating classification industries that, in fact, split.

Table 2: Percent Erroneously Classified, 1982 Summarized at the All-Manufacturing Level

	Reclassification to: 1970 SIC	1980 SIC
4-Digit Level Evaluation		
Product Detail (no resistance)	2.8	2.8
Product Detail (with resistance)	2.5	2.3
Forced One-to-One	1.7	25.6
Mix of Methods	0.8	1.6
3-Digit Level Evaluation		
Product Detail (no resistance)	2.5	2.2
Product Detail (with resistance)	2.3	1.7
Forced One-to-One	1.0	2.9
Mix of Methods	0.7	1.1
2-Digit Level Evaluation		
Product Detail (no resistance)	1.5	1.1
Product Detail (with resistance)	1.3	0.8
Forced One-to-One	0.4	0.4
Mix of Methods	0.5	0.5

IV. A NON-SIC STRATEGY: AGGREGATION

The two main strategies applied in this paper have involved bringing the old standard forward in time and taking the new one back. An alternative is to create a completely new classification by finding aggregations of entire 1970 SIC industries and entire 1980 SIC industries that are equivalent in terms of establishment content. The groupings are then numbered. The result is an aggregation concordance. For any establishment classified on a 1970 SIC or a 1980 SIC basis, comparably-based classification can be achieved by recoding the SIC to the new grouping number.

This strategy of *grouping up* has all the advantages listed for forced one-to-one coding. In addition, no classification error results if this concordance is used for reclassification in the year from which it was generated.

There are three disadvantages: (1) The resulting classes are not as well-known as those of the SIC. (2) There is no simple hierarchical structure. (3) There is loss of detail: the 172 classes of the 1970 SIC and the 239 classes of the 1980 SIC (referred to in the introduction) reduce to just 97 groups--one of which comprises 59 1970 SIC industries and 84 1980 SIC industries.

That loss of detail derives, in part at least, because groupings are generated from actual cross-classified data. This means that unusual production behaviour or erroneous classification can result in additional industries being drawn into a given group. By excluding unusual or questionable inter-industry links in the underlying data, groupings can be prevented from growing in an unwarranted fashion. In this paper, such links are defined to be those in which the overlap between two industries accounts for less than 15% of the value added of each. By excluding those links, a much more detailed concordance has been produced. The result (see reference [1]) comprises 147 industry groupings; no SIC industry is excluded; and no grouping is unduly large. However, excluding any links means that the resulting assignments will be subject to error. That error is equal to the value of establishments whose cross-classification coincides with links deemed unusual or questionable; such error accounts for less than half of one percent of overall manufacturing activity shipments.

A similar sort of concordance is used in the Input-Output tables of the Canadian System of National Accounts. The industry groupings, referred to as link-level industries or historical links, relate 1960, 1970 and 1980 SIC industries. That concordance is not a true aggregation concordance (as defined here) since the groupings do not always comprise *entire* SIC industries. In several cases, SIC industries map to more than one link-level industry. Consequently, reclassification is not always a simple recode of a given SIC industry.

CONCLUSIONS

After testing three methodologies for extending SIC-based classification, the mix of product detail and one-to-one coding was seen to outperform the other methods. It was slightly better when used to extend the 1970 SIC forward in time than when used to take the 1980 SIC back.

There are several relatively minor limitations to the extension of SIC-based classification. The first is that a number of 1970 SIC industries changed in definition while that classification was in effect. This produced breaks in the officially published series that are not a product of this reclassification. These can be handled by reclassifying the underlying data to the 1982 version of the 1970 SIC. A second limitation is that the definition of manufacturing, and therefore the content of the manufacturing industries, changed with the adoption of the 1980 SIC. However, that change was only slight: less than 0.5% of the 1970 SIC version of manufacturing was dropped, and less than 0.5% of the 1980 SIC version is new. A third limitation is that the new commodity classification, an extension of the Harmonized Commodity Description and Coding System, must be linked to the 1970 SIC, before that standard can be extended beyond 1987.

In addition, a number of changes could facilitate future exercises of this sort. First, the resistance rules used in the official series should be codified. Second, all other subjective elements, such as coverage and size significance, should also be codified. Third, a manufacturing services classification should be adopted that is sufficiently detailed to allow unique links to 4-digit industries.

An alternative strategy for achieving historical comparability, and one that is simple and highly accurate, involves the use of an aggregation concordance. By eliminating unusual or questionable inter-industry links in the underlying data, the resultant groupings are kept small and homogeneous. The main disadvantage of this strategy is that the industries are not as widely-recognized as those of the SIC.

In summary, by using a mix of methods to extend SIC-based classification, or by using the non-SIC strategy discussed here, the past twenty years of manufacturing data can be put on a comparable basis of industrial classification.

ACKNOWLEDGEMENTS

Thanks to George Andrusiak (Industry Division), Bruce Cooke (Industry Measures and Analysis Division), Brenda Hutchinson (Industry Division) and Bruce Mitchell (National Accounts and Environment Division) for their very helpful input at a dry run of the Conference presentation. Thanks also to Shaila Nijhowne (Standards Division) for many useful discussions and ideas.

REFERENCES

[1] John S. Crysdale, 'Extending Historical Comparability in Industrial Classification' *Research Paper Series*, Analytical Studies Branch, Statistics Canada, Discussion Paper, forthcoming.

[2] John S. Crysdale, 'Industrial Classification in the Canadian Census of Manufactures: Towards Less Art and More Science' *Statistical Journal of the United Nations Economic Commission for Europe*, December 1988, Volume 5, No 4., 377-392. Also available as 'Industrial Classification in the Canadian Census of Manufactures: Automated Verification Using Product Data' *Research Paper Series*, Analytical Studies Branch, Statistics Canada, Discussion Paper #20, January 1989.

[3] Statistics Canada, *Concepts and definitions of the census of manufactures*, Catalogue 31-528, Ottawa, 1979.

[4] Statistics Canada, *The input-output structure of the Canadian economy, 1961-1981 (Revised data)*, Catalogue 15-510, Ottawa, 1987.

[5] Statistics Canada, 'Notes on the 1980 Standard Industrial Classification in the Manufacturing Industries' in *Manufacturing industries of Canada: national and provincial areas, 1983*, Catalogue 31-203, Ottawa, 1986, xxiii-xcviii.

[6] Statistics Canada, *Standard Industrial Classification Manual, Revised 1970*, Catalogue 12-501, Ottawa, 1970.

[7] Statistics Canada, *Standard Industrial Classification 1980*, Catalogue 12-501E, Ottawa, 1980.

TESTING THE ADVANTAGES OF USING PRODUCT LEVEL DATA TO CREATE LINKAGES ACROSS INDUSTRIAL CODING SYSTEMS[1]

SuZanne Peck, U.S. Bureau of the Census, Center for Economic Studies
Washington, D.C. 20233

Many countries classify business establishments into industrial categories based on information collected about the products or services produced in the establishment. From this micro-level data countries produce aggregate statistics used to measure economic performance. Unfortunately, the classification of economic activity varies over time. To make cross-time comparisons of industrial performance it is necessary to convert the data to a consistent classification system. This conversion may be done at either the aggregate industry level or at the disaggregate product level. Ryten (1991) argues that matching industries at the industry level produces biased measures of economic performance.[2] Ideally, according to Ryten, linking industries should be accomplished by reclassifying product data of each establishment to a standard system, reassigning the primary activity of the establishment, reaggregating the data to the industry level, and then making the desired statistical comparison. The goal of this study is to find what difference, if any, is found in the industry statistics when industries are reclassified using industry as opposed to product level data.

After the major revision in the 1987 U.S. industrial coding system, the Standard Industrial Classification system (SIC), the problem arose of how to evaluate industrial performance over time. The revision resulted in the creation of new industries, the combination of old industries, and the remixing of other industries to better reflect the present U.S. economy. A method had to be developed to make the old and new sets of industries comparable over time. Comparing industries by using the product level data although preferable as noted by Ryten (1991) has two limitations, confidentiality restrictions and expense of using micro data. Using industry level data is simpler. This paper provides two methods for and discusses problems in comparing industrial performance cross-time using both levels of linkage.

To test whether reclassifying establishments using the micro level product data makes an appreciable difference, I convert establishments from the 1982 to the 1987 SIC system. I aggregate total value of shipments (TVS) to the industry level and compare this total to the TVS generated from converting the 1982 industries to the 1987 industries by proportioning the share of the 1982 industry to the 1987 industry. The data used for this project is manufacturing data from the Longitudinal Research Database (LRD). The LRD contains product data by manufacturing establishment, the most micro-level data collected in the U.S.. 1982 is the last year product data are available before the 1987 SIC revision. Using the 1982 LRD file allows the analysis to be made without considering any previous coding changes.

The results are mixed. For the 90 industries unaffected by the SIC revision both conversion methods produce equal industry TVS and since converting at industry level is simpler it is preferred. Linking data at the product level conversion is preferred for 316 industries because it recognizes that establishments may switch their primary industries as a result of the conversion. This results in the two methods' totals differing for one industry. For 53 industries linking data at the industry level produces the most reasonable results due to problems in linking product codes from the 1982 to the 1987 SIC systems. Crysdale (1993) also finds that the best method of constructing comparable industries involves a mixture of methods using both product detail and industry concordances.

Section I describes the industrial classification system in the U.S. and at the Census Bureau. Section II explains the methodology used in converting the establishment level data at the product and industry level. Section III discusses the results of the comparison. Concluding remarks are in section IV.

I. SIC and Census Industrial Coding Systems

In the U.S., the SIC is the main industrial coding system. The basis of the SIC system is the industry, and the basic unit of observation is the establishment or production units. Each establishment is assigned a primary industry code (four-digit SIC code) which represents the primary activity of the unit. The Census Bureau uses the SIC to present most of its economic data. It also samples economic units for surveys based on their industry code. However, for collection purposes, Census expands on the SIC. In order to classify establishments into specific industries Census collects data on products produced in each establishment. Census assigns a seven-digit code for each product that ties the products to particular industries. The first four digits of the product code identify the industry to which the product is primary. In 1987 a major revision of the SIC system took place.[3] Any changes to the SIC system affect Census' coding system. In 1987 Census revised the product codes to account for the new mix of industries.

II. Two Methods of Linking Statistics Classified Under Two Coding Systems

I show two methods of linking data across the 1982 and 1987 SIC codes. Both methods convert the 1982 data to the 1987 system. The first conversion method links the 1982 data to the 1987 data using seven-digit product codes for each establishment, the product code conversion method. The second method involves converting data at the four-digit industry level, the industry code conversion method.

Product Code Conversion Method. Under the product code conversion method, 1982 product codes are converted to 1987 codes using a conversion table. (See 1987 Census of Manufactures and Census of Mineral Industries: Numerical List of Manufactured and Mineral Products, U.S. Department of Commerce, 1989). This table maps products from the 1982 to the 1987 coding system.

Table 1 shows examples of each type of conversion. Example I shows a simple code change where product code 3536257 becomes 3537417. This method reassigns $100 in TVS from the 1982 code to the 1987 code. Example II shows 1982 codes combined into a 1987 code with the $25 and $75 for the 1982 codes assigned to the 1987 code. Example III shows a 1982 code 3573551 splitting into three 1987 codes. It is impossible to proportion out the $100 in TVS from the 1982 code to the 1987 codes since no information exists regarding the distribution of these products within each establishment. In addition, this code change could make assigning the establishment's primary industry code impossible because the largest portion of TVS could go to industry 3572, 3575 or 3577. The product code conversion method arbitrarily assigns the $100 in TVS to product code 3572200. When this represents the largest portion of TVS for an establishment, its primary industry code becomes 3572. This leads to many establishments assigned to some industries and no establishments assigned to ten industries.[4]

The following example summarizes the process of converting 1982 to 1987 codes using the product code conversion method. First, establishments producing the products coded as 3536257 change their product code to 3537417. Next, I aggregate by the first four digits of the product code each establishment's total dollar value of products including the newly coded products. The industry code representing the greatest dollar value becomes the establishment's primary industry code. Finally, I sum statistics for all establishments with the same primary industry code.

Industry code conversion method. The second method, the industry code conversion method, involves converting data at the industry level. First I sum TVS for all establishments in the 1982 LRD to the industry level based on the 1982 coding system. Next, each 1982 industry code is linked to a 1987 code using a linkage table. The linkage table created for the industry conversion method maps codes from the 1982 to the 1987 SIC system by assigning proportions of TVS of the old industry to the new industry. The proportion value represents the percentage of the old industry distributed to the new industry.

Information for the industry level conversion table comes from two sources. The first source of information, Appendix A - Section III in Standard Industrial Classification Manual 1987 (Executive Office of the President, 1987) shows the relation of the 1977 to the 1987 SIC industries and serves as the foundation for creating the conversion table. The table provides all industry code changes that occurred in the 1987 SIC revision. Table 1c-2 of the 1987 Census of Manufactures Industry Series reports (U.S. Department of Commerce, 1990) serves as the second source of information. This source provides the means to calculate the proportion of TVS of the old industry assigned to the new industry. During the processing of the 1987 Census of Manufactures, Census computes for each establishment a 1982 and a 1987 primary industry code. Table 1c-2 shows the TVS for 1987 tabulated by the 1982 industry code. Each 1982 industry has a corresponding 1987 industry and a proportion of TVS for the 1982 industry that went into the 1987 industry. From this table, I calculate the proportion of the 1982 industry converted to each 1987 industry.

Table 2 clarifies the method of proportioning out the TVS for 1982 industries to 1987 industries. From the 1987 SIC Manual, I know which industries did not change or simply changed industry code between 1982 and 1987. For these industries to convert the 1982 industry data to the 1987 system I multiply the industry TVS by one. In example I industry 3331 does not change between 1982 and 1987. To convert industry 3331 I multiply the 1982 TVS in 3331 by one coming up with the 1987 TVS for industry 3331. From the 1987 SIC Manual, I also know the 1982 industries combined in 1987. These industries are also assigned a proportion of one. For example, as shown in example II, TVS for industry 2351 and 2352 are both multiplied by one and reassigned to industry 2353.

Industries rearranged or split into new industries require information provided in table 1c-2 to proportion out data from the 1982 industry into the 1987 industries. Example III shows how this is done for industries rearranged in 1987. Industries 2047 and 2048 exist in both years of this analysis, however, products shifting from industry 2047 to industry 2048 make the two

industries look different in the 1987 system from the previous system. Multiplying TVS for industry 2047 by 0.9 gets the 1987 industry value for industry 2047 and by 0.1 gets the portion of industry 2047 now in industry 2048. The TVS for industry 2048 is reassigned to the 1987 industry 2048.

Example IV shows the TVS for 1982 industries proportioned out to new 1987 industries. From table 1c-2, I calculate the proportion of industry 2831 distributed to the 1987 industries 2835 and 2836. I multiply the TVS for industry 2831 by 0.58 to get the TVS assigned to industry 2835 and by 0.42 to get the TVS assigned to industry 2836.

The summary of the process of converting 1982 to 1987 codes using the industry conversion method is as follows. Industry level data summed from the LRD and the conversion table are merged. I multiply the TVS assigned to each 1982 industry code by the number in the proportion column. The result is the value assigned to the 1987 code. The dataset created contains 1982 data summed to industries that resemble 1987 industries.

III. Comparison of Product Code and Industry Code Conversion Methods

The product and industry code conversion methods link 1982 industrial codes to 1987 codes. In this section, I calculate the percent difference in TVS for 1987 industries created from the two methods of conversion. To illustrate the difference in the totals generated under both conversion methods, I graph the intersection of the logarithm of the two values. The more distant the intersection is from the 45-degree line the greater the difference in the two methods. Discussion of the source of these differences follows.

To simplify the analysis, I separate the 459 1987 industries into four groups. Table 2 shows the group definitions. Unchanged industries from the 1982 to 1987 coding systems make up group I. Group II consists of 1982 industries combined into one 1987 industry. Group III represents industries that remained in the 1987 system, however, no longer resemble the former industry. For group IV, I include all 1987 industries that resulted from old industries splitting and creating new industries.

Measuring the Difference in Levels. In this section I show the difference in the 1982 statistics generated from the two conversion methods described in section II. To compare the two methods I measure the percent difference in TVS for each 1987 industry. The percent difference measure is

$$d_i = \frac{|y_i - x_i|}{x_i} * (100) \qquad (1)$$

where y_i is TVS for industry i for 1982 from the product code conversion method and x_i is the value for industry i for 1982 from the industry code conversion method. Industry i represents 1987 codes. Finally, d_i represents the percent difference in TVS for the two methods of conversion for one industry.

From the analysis, I expect d_i to be smaller for groups I and II than for groups III and IV because groups III and IV experience more change in their industry definitions than groups I and II. Table 3 presents the mean and maximum values of d_i for each group of industries. The values of d_i range from zero to 625 percent. For ten industries d_i equals 1 because y_i equals 0. This occurs for ten industries that had no establishments classified to them under the product code conversion method.

Figures 1a-d plot $\ln(x_i)$ on the x-axis and $\ln(y_i)$ on the y-axis to show the differences in the two methods. Points that lie along the 45-degree line represent industries where the methods do not differ. The farther a point is from the line the greater the difference in the conversion methods. The ten industries in group IV where y_i equals zero are missing from figure IV.

As expected, for groups I and II very little difference exists between the two conversion methods. For group II the mean d_i is less than 1 percent. For group I the mean d_i is 2.7 percent. Throwing out the 84 industries where d_i equals 0 raises the mean d_i to 3.6 percent. Figures 1a and 1b show that for group I and II industries very little difference exists in the TVS generated under the product code and industry code conversion methods.

Two sources of the difference between the two methods in TVS in these industries are product codes within the industry changing and the establishment's product mix changing and redefining its primary industry. Under the product code conversion method an establishment may switch its primary industry code due to a shift in its product structure. In industries where plants produce a large volume of secondary products, the two conversion methods could produce different results. The product code conversion method accounts for the movement of secondary products while the industry code conversion method does not. The mean of d_i is larger in groups III and IV because industries in these groups experienced more coding change between 1982 and 1987. In table 3 the average d_i for group III is 8.7 percent and for group IV is 66.0 percent. Ex-

cluding industries where d_i equals zero raises the mean for group III to 9.2 percent. For group IV, after subtracting the ten industries where d_i equals one or y_i equals zero the mean d_i falls to 57.9 percent.

Figure 1c shows that for group III more points lie off the 45-degree line than for group I. The mean d_i is 6.0 percent larger for group III then for group I. For group III the product code conversion method captures switches in establishments while the industry code conversion method does not. The point for industry 3679, electronic components, not elsewhere classified, lies off the 45-degree line. In the 1987 SIC revision, five products shifted from this industry to five industries. Most of the changes for industry 3679 involved conversions like examples I and II in table 1. However, one product code was distributed to two new product codes in different industries. The mean of d_i for these two methods differs due to the handling of the product structure change.

Obviously from figure 1d, the method of conversion makes a difference on newly created industries. In this graph many points lie off the 45-degree line. The methods differ on average by 57.9 percent after eliminating the ten industries where under the product code conversion method their TVS equals zero. For some industries these methods produce large differences due to the lack of product detail in 1982. Example III in table 1 illustrates one example. Another example is the electronic computing equipment industry, 3573, which is broken into five new industries in 1987: 3571, 3572, 3575, 3577, and 3695. Here, product data for the five new industries were not collected under the old coding system in enough detail to break out the products to the new industries. In figure 1d points for industries 3575 and 3572 lie off the 45-degree line and industries 3577 and 3695 do not appear on the graph at all. The product code conversion method allowed more establishments assigned to industries 3575 and 3572 and none assigned to industries 3577 and 3695.

IV. Conclusions

The goal of this study is to find what difference if any is found in the industry statistics when industries are reclassified using industry as opposed to product level data. When creating a time series of industry data between 1982 and 1987 when a major coding change occurs, I conclude that the preferred method depends on the industry.

For 90 industries from groups I, II, and III no difference exists between the two conversion methods. Although, both methods produce similar results when converting industries the preferable method is the simpler industry code conversion method. For 316 industries, 69 percent of all 1987 industries, the product code conversion method is the best method of conversion. This method recognizes establishments switching industries so it better represents real changes in the industries. For the remaining 12 percent of 1987 industries I recommend using the industry code conversion method. For these 53 industries the industry code conversion produces the most reasonable results. Linking emerging industries over time at the product level is difficult and in some case impossible because new products did not exist in the earlier years.

In order to link industries the method of reclassifying industrial data over time matters. Even in linking codes across two vintages of the same coding system the method produces different results. Usually the product code conversion method works best. For new emerging industries linkage across time is very sloppy so it is better to use the simpler industry code conversion method.

REFERENCES

Crysdale, J. S. 1993. "Extending Historical Comparability in Industrial Classification." presented at International Conference on Establishment Surveys, Buffalo, NY, June 27-30, 1993.

Executive Office of the President, Office of Management and Budget. 1987. 1987 Standard Industrial Classification Manual, Government Printing Office, Washington, D.C..

Ryten, J. (1991). "Inter-Country Comparisons of Industry Statistics," in 1991 International Conference on the Classification of Economic Activities Proceedings, November 6-8, .

U.S. Department of Commerce, Bureau of the Census. (1989). 1987 Census of Manufactures and Census of Mineral Industries: Numerical List of Manufactured and Mineral Products. [Washington, D.C.].

_____. (1990). 1987 Census of Manufactures: Industry Series. [Washington, D.C.]..

_____. (1991). 1987 Census of Manufactures: General Summary. [Washington, D.C.].

ENDNOTES

1. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau. I benefitted from the comments and assistance of Robert McGuckin , Harvey Monk, Kenneth Troske, and Lynn Weidman.

2. Linking industries over time is problematic because establishments produce more than one product or provide more than one service. When linking industries one assumes homogeneous establishments in an industry. Using commodity data (five-digit product class data) alleviates some of this linkage problem. (Ryten, 1991).

3. The previous major revision occurred in 1972. The last year industry codes changed was in 1977. No industry codes changed in 1982.

4. The ten industries are: 2835, 3082, 3084, 3085, 3088, 3492, 3577, 3593, 3672, and 3695.

TABLE 1
SAMPLE OF THE TABLE FOR THE CONVERSION OF PRODUCT CODES*

	1982 product code	TVS**	1987 product code	TVS
I.One to One Comparison	3536257	$100	3537417	$100
II.1982 Codes Combined in 1987	3545143	$75	3545153	$100
	3545144	$25		
III.1982 Code Splits in 1987	3573551	$100	3572200***	$100
			3577200	$0
			3575200	$0

* U.S. Department of Commerce, 1989
** The column labeled TVS (total value of shipments) shows how $100 in TVS is distributed under the each coding system.
*** Data classified under these product codes ending in '00' are collected under the Current Industrial Reports program.

TABLE 2
A SAMPLE OF THE TABLE OF PROPORTIONS OF 1982 INDUSTRIES IN 1987 INDUSTRIES

	1982* industry	1987* industry	proportion**
I. No Change	3331	3331	1.00
II. 1982 Codes Combined in 1987	2351	2353	1.00
	2352	2353	1.00
III. 1982 Codes Rearranged in 1987	2047	2047	0.90
	2047	2048	0.10
	2048	2048	1.00
IV. 1982 Code Splits in 1987	2831	2835	0.58
		2836	0.42

* The industry columns are created from the table in the 1987 SIC manual showing the relation of the 1977 to the 1987 SIC industries (Executive Office of the President, 1987).
** The proportions are calculated from information in table 1c-2 in the 1987 Census of Manufactures: Industry Series (U.S. Department of Commerce, 1990).

TABLE 3
PERCENT DIFFERENCE BETWEEN X_i AND Y_i

	Number of 1987 Industries	Mean percent difference (d_i)	Maximum percent difference (d_i)
I. No Change			
a. All industries	341	2.7 %	39.4 %
b. Exclude industries where $d_i = 0$	257	3.6 %	
II. 1982 Codes Combined in 1987			
a. All industries	4	0.4 %	1.0 %
b. Exclude industries where $d_i = 0$	2	0.9 %	
III. 1982 Codes Rearranged in 1987			
a. All industries	61	8.7 %	59.2 %
b. Exclude industries where $d_i = 0$	57	9.2 %	
IV. 1982 Code Splits in 1987			
a. All Industries	53	66.0 %	624 %
b. Exclude industries where $d_i = 1$	43	57.9 %	

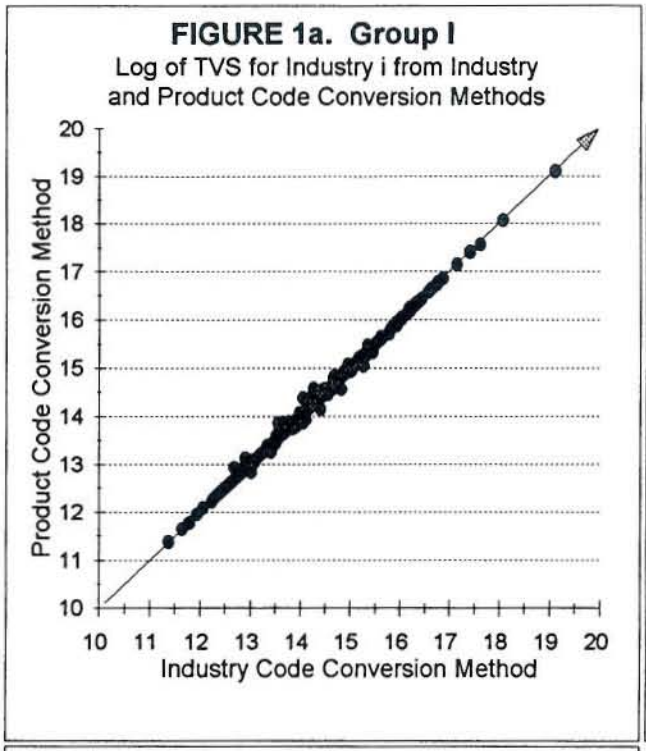
FIGURE 1a. Group I
Log of TVS for Industry i from Industry and Product Code Conversion Methods
Product Code Conversion Method
Industry Code Conversion Method
10 11 12 13 14 15 16 17 18 19 20

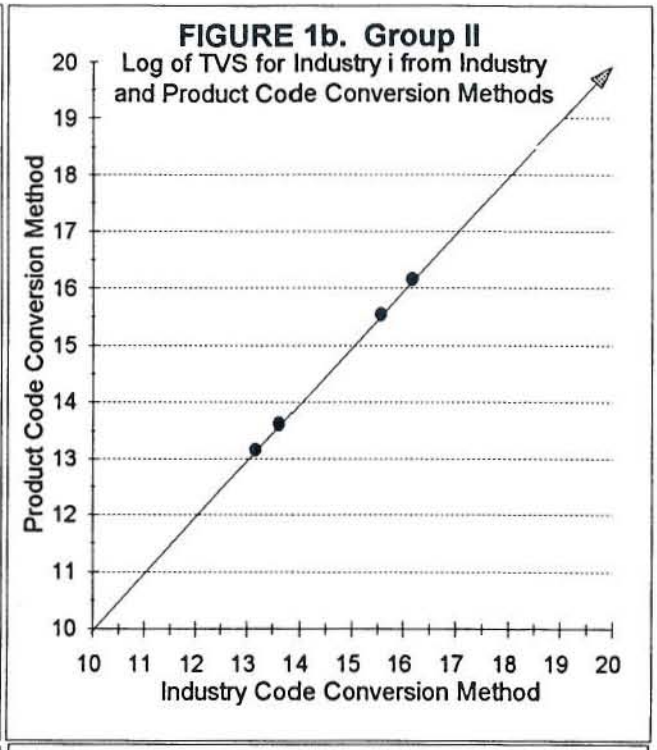
FIGURE 1b. Group II
Log of TVS for Industry i from Industry and Product Code Conversion Methods
Product Code Conversion Method
Industry Code Conversion Method
10 11 12 13 14 15 16 17 18 19 20

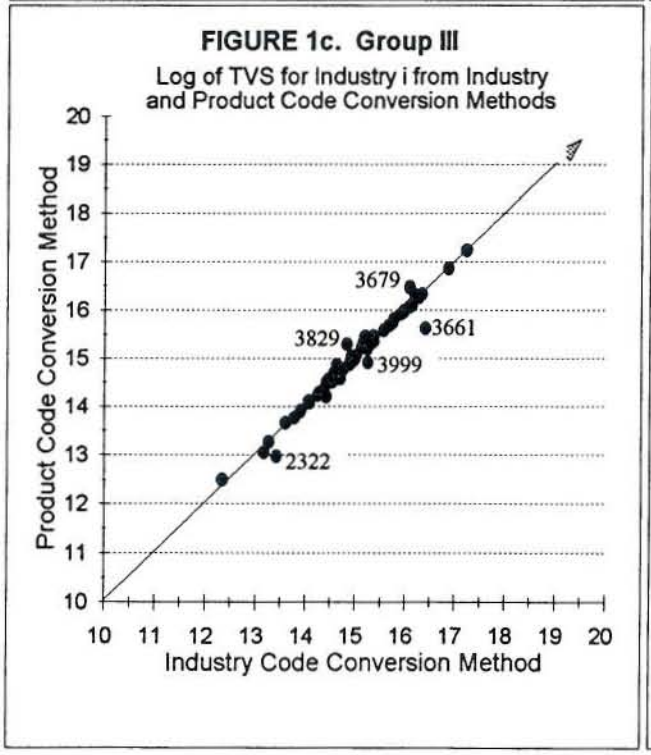
FIGURE 1c. Group III
Log of TVS for Industry i from Industry and Product Code Conversion Methods
Product Code Conversion Method
Industry Code Conversion Method
10 11 12 13 14 15 16 17 18 19 20
3679
3661
3829
3999
2322

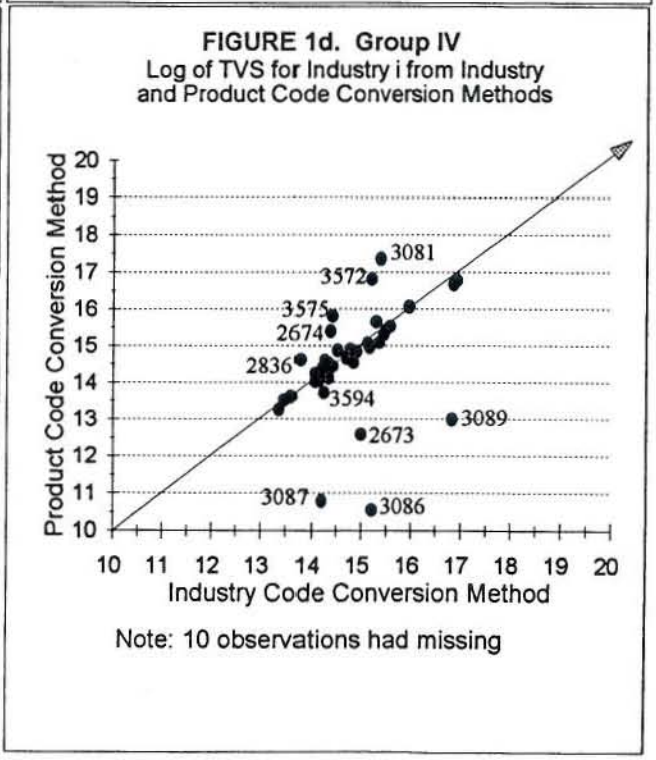
FIGURE 1d. Group IV
Log of TVS for Industry i from Industry and Product Code Conversion Methods
Product Code Conversion Method
Industry Code Conversion Method
10 11 12 13 14 15 16 17 18 19 20
3081
3572
3575
2674
2836
3594
3089
2673
3087
3086
Note: 10 observations had missing

The ARIES Review System in the BLS Current Employment Statistics Program

Richard Esposito, Dong-yow Lin, and Kevin Tidemann, Bureau of Labor Statistics
Rich Esposito, Room 5890 PSB, 2 Mass. Ave. NE, Washington, DC. 20212

KEY WORDS: Graphics, Editing, Statistical, Survey

Recent advances in PC computing power have made it possible to replace the arduous computer listing-intensive review of estimates and sample data in large surveys with easier to use graphical and query methods. Such a graphical and query system has been successively used in the Current Employment Statistics (CES) program of the Bureau of Labor Statistics of the U.S. Department of Labor since November, 1991. This new ARIES (Automated Review of Industry Employment Statistics) system has improved the data review capabilities of CES analysts, and has occasioned a fundamental rethinking about the possibilities for improving data review in the BLS. ARIES, as so far developed, uses a top-down approach for outlier detection and treatment, in that the search for suspect sample data is driven by preliminary identification of suspect estimates. Graphical depiction of industry estimates which deviate from historical trends enables viewers to limit their search for sample outliers to a small subset of the numerous industry estimates computed, and query and graphical search techniques are then used to pinpoint individual suspect sample members which may have caused those estimates to be suspect. Finally, outliers are given appropriate weights and new estimates calculated using easy to use tabular screens on the PC.

Description of the Current Employment Statistics Program

The CES program of the BLS estimates total employment, employment of women, non-supervisory employment, average weekly hours, and average hourly earnings for virtually all non-agricultural industries in the United States, and average overtime hours in manufacturing industries. These estimates are computed from a month-to-month matched sample survey of over 300,000 governmental and private establishments each month, with estimates computed for over 1600 sample-based and 1000 aggregate level industry cells, for each of the data elements mentioned. The quality of the sample data and estimates is the responsibility of about a dozen industry analysts who guide the data through a series of quality control steps, from the receipt of the sample data from state offices which collect the data from establishments, through the process of computing final estimates.

Description of the Pre-ARIES CES Sample Review Process

For about 16 years, data review in the CES program has been a mix of powerful main frame computers utilized for number-crunching, and powerful human frame computers utilized for paper-crunching. Both machine and human computers can be said to have been underutilized. Many tasks which would have made the human task less arduous were never assigned to the machine, and humans and machines were not involved in feeding the other side useful information which would have made both sides' tasks more efficient. This is probably the point on which ARIES makes its most significant contribution. The information which we intend the ARIES computer system to provide is essentially a better understanding of the data on a current basis. The old system of reviewing many thousands of lines of computer paper owed its success to the ability of the human industry analysts to construct over time their own internal pictures of the industries, gaining, through arduous labor and from bits and pieces, some idea of the whole.

The ARIES System: A Short Description

Industry employment and earnings estimates are computed from establishment sample data 3 times each month; first preliminary, 2nd preliminary, and final estimates are computed, each estimate using more cumulated sample as establishments report their data for the given month. Typically for 1st preliminary estimates, 200,000 sample data reports are processed on an IBM mainframe. For all industry estimating cells which fail tolerance checks, all sample data and associated historical data are downloaded from the mainframe to 12 individual 386-based PC's, according to the industries for which each of 12 industry analysts is responsible. This transfer of data from the mainframe to PC's is done automatically overnight, using a combination of programs written in C and ACS Excellink/Host-V. An investigator program automatically notifies the mainframe computer to begin alternative fall-back procedures if any individual PC fails to receive its data.

By the time industry analysts arrive at work in the morning, all relevant data has been automatically downloaded to their individual PC's and the review

procedure can begin. A short description of this ARIES review process is as follows: each analyst uses an estimate level graphical representation of all of their industries to identify those industries with suspect estimates, based on normal historical month-to-month changes. For each suspect industry estimate identified, the analyst will use either of two graphical methods, or a query method, to identify suspect sample reports which may have contributed to the suspect estimates. Suspicious sample thus identified can then be automatically corrected or given an appropriate weight for estimation, and new estimates are automatically and immediately calculated based upon any weightings performed.

All changes made are automatically entered into an audit trail, and also copied to a LAN server, so that a permanent record is kept for supervisory review and later analysis. At the end of the review process, corrected and weighted sample data are uploaded back to the mainframe for storage.

Development Considerations in ARIES

Each step in ARIES has not only fulfilled a specific need, but has also generated ideas for future improvements.

Industry analysts begin their review with the "Anomaly Map" for their own specific industries.

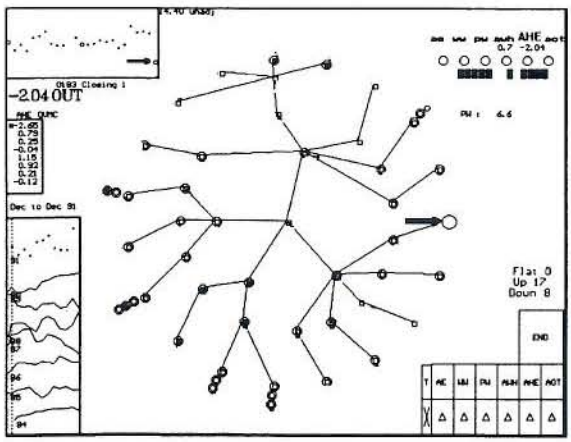

Figure 1: Anomaly Map

A typical anomaly map is shown in Figure 1, with all industry identifying information eliminated because of confidentiality requirements.

The anomaly map is a tree of the industry structure for that industry analyst, with more finely differentiated estimating cells on the perimeter, and more aggregate levels closer to the center of the map. Essentially, each node of the tree represents a specific industry. Estimation from sample data is done for perimeter cell industries, and estimates for more aggregate level industries towards the center are obtained by addition from the perimeter cells. On the PC, colors are used to mark industry nodes whose estimates are outside historically determined tolerances, and the specific color used indicates how much out of tolerance that industry's estimate is. The connecting lines of the tree are also colored to connect parent and child nodes which have similar tolerance failures. When any given industry node is clicked on with the mouse, historical monthly estimates are charted stacked-by-year in the lower left corner, and for the most recent two years in the upper left corner. In Figure 1, historical time-series are shown for the industry marked with the larger arrow (for this article, a large size circle has been substituted for the color indicator on the PC). The smaller arrow in the upper left corner points to the current estimate, which can clearly be seen to be far below the historical trend for this industry. The anomaly map shown here shows tolerance anomalies for Average Hourly Earnings estimates; by using the mouse, the colors will change to show tolerance exceptions for Employment, Woman Worker, Production Worker, Average Weekly Hours, or Average Overtime Hours.

The anomaly map was created for two reasons. The first motivation was to give industry analysts a better picture of what is happening to the estimates in their industries, and the second was to enable a quick top-down search for industries whose sample may be causing an estimate to be suspect. Suspicious estimates at the published level can quickly be traced back to the specific basic estimating node which caused the deviation. The design impulse to create the anomaly map was to try to fit as much information on all of the industries on a single screen. The anomaly maps can be seen as a first step in obtaining an overall picture of what is happening in a large subgroup of industries. At present, only one category of estimates is shown at a time, for example, in Figure 1, only the colors for average hourly earnings would be plotted. Ideally, the relationships between all categories for all industries could be pictured at once. We supply a partial solution to that problem by plotting the color representation, in the six circles in the upper right corner, for all the associated data categories for a single industry, once that industry's node has been pointed to by the mouse.

That eliminates a lot of paging back and forth between screens.

The anomaly map is used to concentrate the search for suspect sample data on only those industries which have problems with their estimates. Once an industry analyst has identified those industries on the anomaly map, he or she will begin the search for suspect *sample* data, confining the search to only those suspect industries.

To isolate suspect sample data in a given industry, two graphical methods and one query method are available. The first graphical method is to identify sample outliers from a scatter gram such as Figure 2.

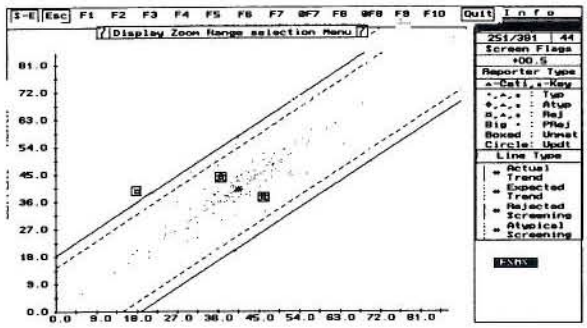

Figure 2: Scatter Gram for a Single Industry

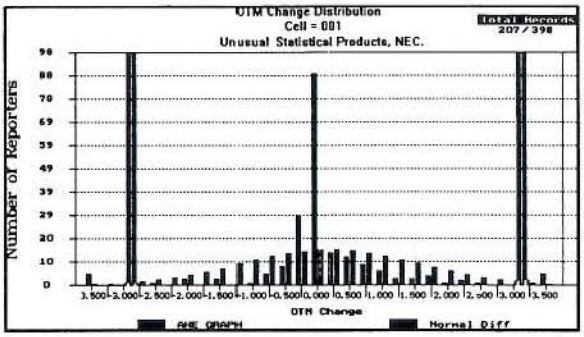

Figure 3: Sample Distribution for a Single Industry

Each industry's scatter gram takes a few seconds to plot on the PC screen. Each point on the scatter gram represents the current month's reported data (on the vertical axis) and the previous month's reported data (on the horizontal axis) for an individual establishment. In the CES survey, month-to-month matched sample data is used to calculate estimates. A natural expectation is that data points will group along the 45 degree line, with seasonality and trend factored in. Employment estimates tend to group along this line better than average hourly earnings, average weekly hours and average weekly overtime hours worked. Automatically computed screening ranges are constructed and plotted on the screen at a set number of standard deviations from the (approximately) 45 degree expected value line. The establishment points outside the solid lines (extreme failure region) and those outside the dashed lines (less extreme failure region) would normally be the establishments whose data the industry analyst would select, using the mouse, to look at more carefully. The information boxes to the right in Figure 2 contain a legend for the various actions performed on each point, and these actions and various categories of sample data are marked with special symbols on the scatter gram. By selecting one or more scatter gram points with the mouse, the analyst can quickly bring up on the screen detailed tabular and graphical historical sample data corresponding to the establishment represented by that point.

The second graphical method used to identify sample outliers is through the plotting of the sample distribution of the month-to-month change in all sample establishments in an industry. An example of such a distribution graph is seen in Figure 3.

Selection of sample outliers to further investigate is done from the distribution graph by moving the tall double bars to isolate sample reports on either end of the distribution. The sample distribution itself is shown as the darker bars; a comparison normal distribution is also plotted as the less dark bars.

The third method used to isolate sample outliers is through the use of pre-set query questions; various parameters can be set for any of 108 different queries, and the PC will search the particular industries sample data to select those sample members which meet the conditions set by the queries. The query mode of selecting outliers is especially suitable for finding sample with specific characteristics, or with particular relationships among their different sample data items.

Once outliers are isolated using any of these three methods, the industry analysts will give suspect sample data appropriate weights, and new estimates are immediately calculated on the PC and added to more aggregate industry levels.

Recent Prototype Developments

In order to provide a better and more comprehensive visual representation of data, we are attempting to provide solutions to several specific goals which the ARIES system has brought into focus. For this purpose we have constructed a prototype system which we anticipate will be eventually integrated into the ARIES system. This "Phase 2" prototype is seen in Figure 4.

In Figure 4, the scatter grams for all six sample data types for all sample reports for an industry are shown on one screen, so that each sample establishment is represented by six points, one in each scatter gram. When a single point in any of the six scatter grams is clicked using the mouse, the associated points in the other five scatter grams for that establishment are circled, as well. When more than one establishment point is captured by the mouse, associated circles are color coded to keep the establishment information

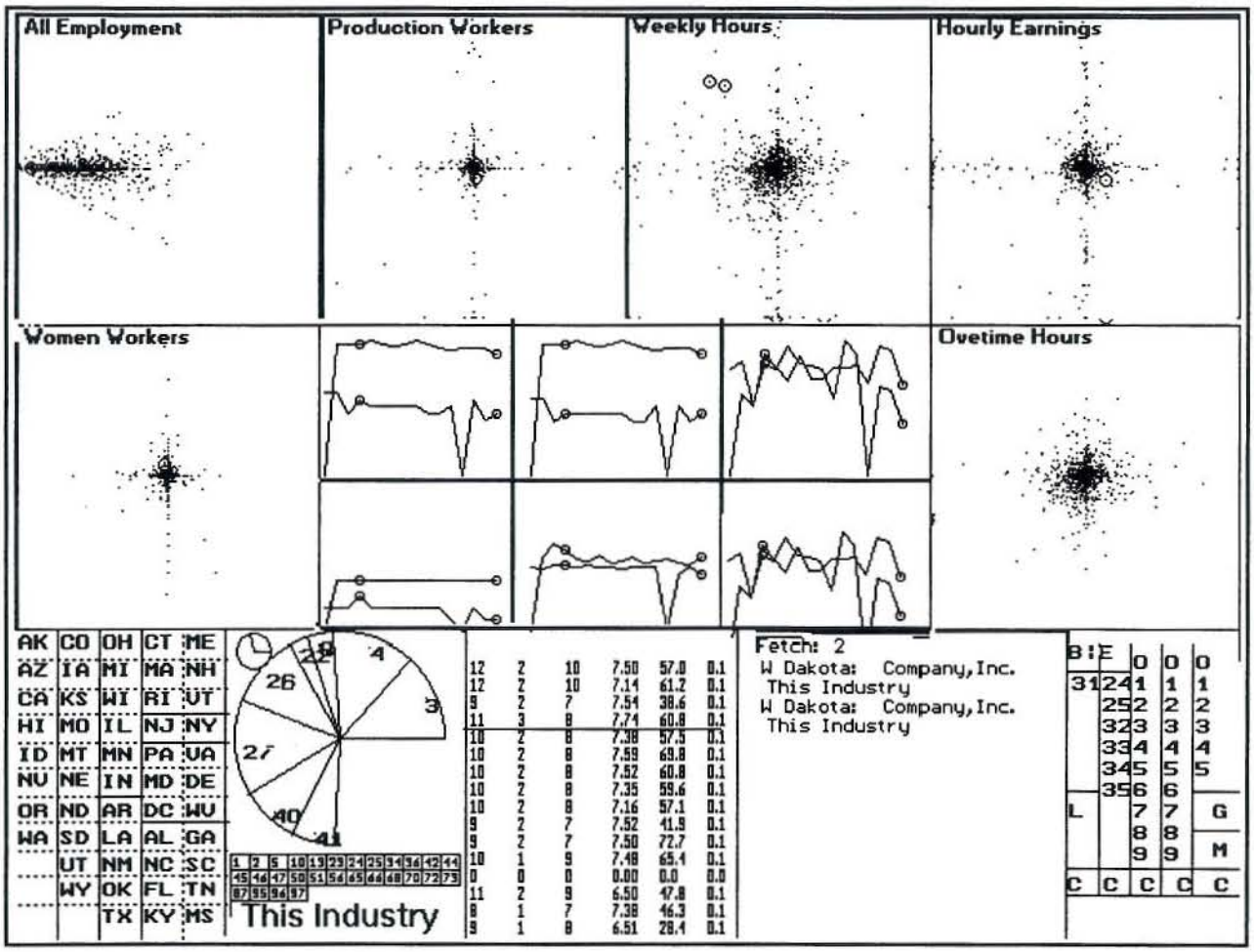

Figure 4: Interactive Scatter Grams for All Data Types and Information by State, Comment Code and Firm

separate. At the same time, the establishment name and address and other identifying information appear in the bottom row, fourth box from the left (replaced here by company names from West Dakota), and a graphical time-series for the most recent 16 months of sample data are displayed for all six data types in the smaller boxes in the center of the screen (also color coded on the PC screen). In the bottom row, 3rd box from the left in Figure 4, is shown the numerical values for those 16 months for any of the establishments shown, which appear by mouse-clicking on the company name. In Figure 4, we have mouse-clicked the points for two sample member establishments.

These scatter grams are an evolution from the scatter gram of Figure 2. In Figure 2, only one sample data type is pictured on the screen at one time; in Figure 4, all data types are shown, and this makes it much easier to see relationships among data types for the same establishments. For our review, that is very important. In Figure 4, the horizontal axis represents the change from the previous month to the current month in the current year, and the vertical axis represents the change from the previous month to the current month's value lagged one year (An exception is the All Employees scatter gram). Thus, points along the 45 degree line represent establishments whose data activity this year matched their own data activity last year. Zero month-to-month change this year and zero month-to-month change last year is represented in the very center of each scatter gram. As an option, the horizontal axis here for the All Employment scatter gram represents the actual magnitude of the establishment, rather than the change from last month, in short, the size of the firm. In this case, the vertical axis has zero point at the midpoint of the vertical axis, and represents the difference in month-to-month movement between this year and last year. Thus, when any point is selected by the mouse, the size of that establishment's employment is instantly obvious.

How would an industry analyst use Figure 4 on his or her PC? Selection of the industries for display of scatter grams is controlled by the selection box in the lower right hand corner. The number code of an industry is selected by the mouse, and the analyst can choose to cycle through a large group of industries one at a time, or view combined industries. The scatter grams can be limited to show only establishments of a certain size, which would be important in finding only the most significant sample outliers. Once the selection of an industry to view has been made, the scatter grams for that industry appear. At the same time, each state in the state map changes color to indicate an index of sample activity for that state. One can then see if an estimate level problem is limited to a single state or subgroup of states. The state map can also be used to plot scatter grams for only a particular state.

The pie chart to the right of the state map is a pie chart of the percentage of specific explanation codes used when reporting sample data. For example, a state agency will affix a "strike" code if sample values for an establishment have been affected by a strike. The pie chart thus shows at a glance some of the most important factors influencing the data in the current month. Codes which individually comprise less than two percent of all comment codes received are arranged as tiny boxes below the pie chart.

When the scatter grams first appear on the screen, the center six time-series boxes are reserved to describe and portray information about sample data and estimates. These machine descriptions will be constructed from automatically prioritized judgements based on historical and cross-sectional characteristics. This is still in the development phase.

Conclusion

We view the successful introduction of the ARIES system as just a beginning. The principles that we have followed and have been led to can be said to be the following:

1. Make the system applicable to the task at hand.
2. Try to make information digestible by using graphics.
3. Make the system interactive, to increase flexibility of use.
4. Make a system for which it is more natural to make improvements rather than a system for which it is more natural to just use and ignore. This means, make a system which gives information which is useful to guiding future improvements.

A STATISTICAL INVESTIGATION OF FARM ACCOUNTING DATA NETWORK

Carlo Filippucci, Università di Bologna Fabrizio Alboni, Alberto Fabbiani, RES coop, via Belle Arti 41, Bologna, via S. Vitale 56, Bologna, Italy

KEY WORDS: Repeated surveys, restricted GLS, post-stratification, farm accounting

" The advantage of randomization is that if a randomized design has been employed no further justification is needed; the whole scientific community will accept the sample that has been selected. With other forms of sampling users would need to be convinced in each case that the sampling scheme could be ignored." SMITH (1983)

1. Introduction *

This paper deals with an attempt at exploiting an important source of data which were collected in order to investigate farms belonging to the European Economic Community (EEC): the Farm Accounting Data Network (FADN). This source, was created in 1965 by European Community and has been implemented in Italy from 1967 to meet the Community Agricultural Policy information needs.

The goals of FADN are to create a data base for the agricultural sector and to define a common methodology for data collection to set up a comprehensive information system for observing and analyzing agriculture both in the whole European Community and in its Regions. Such analysis should be carried out at macro and micro level. For this reason the information to be collected concerns: i) structural data about farms, i.e. land use, labour, equipment, indebtedness, etc.; ii) detailed data for all main livestock and crop productions, i.e. data concerning a detailed description of costs and returns.

From the above description, FADN seems to be an important source of data for the agricultural economic policy because it contains a great amount of information and gives the possibility to obtain macro and micro data about the various countries and regions of Europe. In fact, many studies, based on FADN, have been carried out (among them, we limit to recall the Dubgaard, Grassmugg, Munk 1984).

In spite of the effort of EEC to define a common methodology for data collection, different criteria are followed by each country. In some cases a sample is selected but, in some others, neither a sample design is arranged nor random selection of farms is performed. The heavy burden of survey for sample units, due to the complexity of keeping detailed records as requested by FADN, could explain the difficulties met in implementing a proper sample survey.

The FADN is even more interesting for Italy, where a national survey aimed at obtaining data on structural aspects of farms as well as economic data is not carried out. Moreover, in Italy, agricultural GNP is not estimated from farm data and Agricultural Censuses gives only structural data. Italy contributes to FADN by collecting data from a large number of farms in each region but the survey cannot be considered a proper sample survey. In fact, in Italy the selection of participants is determined by two main devices: i) the farms asking for financial support from the ECC or from Regions are requested to keep their accounts according to FADN; ii) the farms asking for assistance in improving farm management are included in FADN. For above reasons and because of the absence of any sample design the survey can be considered a sort of administrative source.

Another aspect of FADN must be stressed: because of the particular reasons leading farms to participate in FADN, it is possible to observe a portion of farms at more than one time point (*repeated survey*). The interest in such a survey is considerable and increasing (basic references are: Duncan, Kalton, 1987; Kasprzyck, Duncan, Kalton, Singh, 1989; Fuller, 1990; Statistics Canada, 1992), in fact, the measure of changes and of individual behaviours is the increasing need for economic and social sciences as well as for government and official statistics. Furthermore, as it is well known (Duncan and Kalton, 1987), repeated observations of sample units improve measures of net change, to reduce variance of the estimates at any specific time, to measure components of change: gross change, change and variability for an individual (important applications are for micro econometric modelling) and to obtain information to study some important aspects of data quality.

The interest for the information provided by FADN, the lack of the official sources about data for agriculture, especially with repeated observations, and the large resources devoted to FADN were the main reasons that led us to undertake a statistical investigation aimed at exploring if FADN can be exploited as a statistical source. The study started by referring to one Italian region very important for agriculture, Emilia Romagna. In particular, this paper deals with two main aspects: i) to introduce some considerations and analyses aimed at clarifying how and if data could be used; ii) by taking into account the longitudinal structure of data, to apply some proper

* Research supported by ERSA-Italy

techniques to obtain estimates exploiting the information given by overlapping as much as possible.

We feel that the work done for FADN might be useful for other similar administrative sources.

2. Some characteristics of FADN in Emilia Romagna

The National Institute of Agricultural Economics (INEA) is responsible for FADN but the Farmers Associations are entrusted with data collection.

The FADN's universe includes no farms having less than 2 ESU (European Size Unit, about $ 2,800 of Standard Gross margin). Below such a threshold it is very difficult to identify a real agricultural activity: about this question, we refers to "part-time" farms [1].

As we have previously stated, a proper survey design does not exist and only some general guidelines are given by INEA: as a consequence, in many regions (as in Emilia Romagna) more farms than requested are observed[2]. In Emilia Romagna during the last 5 years the percentage of farms taking part to FADN to have assistance has increased to 70% of the total and the 30% only partecipates because of a request of financing. This result is quite important. In fact, participants in the survey because of a request for financing might not be willing to keep records, considering the survey as just a bureaucratic burden; moreover they may be considered a biased set of the farm population, that is the farms more prone to innovate. On the contrary, the request for assistance comes from a more generalized need of farms to face deep agricultural changes, so participation cannot be ascribed to a particular group of farms.

Interviewers are agricultural technicians having a basic knowledge of agriculture and accounting but poor statistical skill. From a statistical point of view, the main problem to face is *no random selection* of farms[3].

The investigation of the survey in Emilia Romagna has been limited to the period 1986-1991 (on the average about 1750 farms), both because the above mentioned reasons for participation and the increasing quality of records that we have checked in the last few years [4]. About the overlapping of farms taking part in FADN (1986-1991) see tab 1.

TAB.1 - Overlapping patterns over time in Emilia Romagna

86	87	88	89	90	91	period of permanence	n° farms	%
X	X	X	X	X	X	6	702	21.0
X	X	X	X	X	O	5	162	4.9
O	X	X	X	X	X	5	199	6.0
X	X	X	X	O	O	4	118	3.5
O	X	X	X	X	O	4	198	5.9
O	O	X	X	X	X	4	56	1.7
X	X	X	O	O	O	3	122	3.7
O	X	X	X	O	O	3	24	0.7
O	O	X	X	X	O	3	15	4.5
O	O	O	X	X	X	3	20	0.6
X	X	O	O	O	O	2	562	16.8
O	X	X	O	O	O	2	15	0.5
O	O	X	X	O	O	2	30	0.9
O	O	O	X	X	O	2	82	2.5
O	O	O	O	X	X	2	44	1.3
X	O	O	O	O	O	1	450	13.5
O	X	O	O	O	O	1	137	4.1
O	O	X	O	O	O	1	20	0.6
O	O	O	X	O	O	1	49	1.5
O	O	O	O	X	O	1	92	2.8
O	O	O	O	O	X	1	12	0.4
X	O	X	X	X	X	5	20	0.6
X	X	O	X	X	X	5	16	0.5
others patterns							60	1.8
TOTAL							3341	100

3. Could FADN be considered representative of farm population?

As we have seen, sample representativeness is the main problem to face. We could choice to limit ourselves to the observed subset of farms leaving data users free to decide how and if it is possible to generalize the results. This is not a solution and, it leaves room for unsuitable and uncontrolled generalizations. Moreover, both the potentialities of the survey and the resources employed were inadequately exploited. While a new and proper survey would be a better solution to meet the need of data asked by policy makers and farmers, the present lack of public resources proves prohibitive. For these reasons using FADN data seemed to us worth to investigating.

Our problem could be faced in two different ways. One involves reasoning on the meaning of *representative sample*; the other is looking for possible *ignorability of non-random sample selection.*

---

1 A great deal of literature exist on this topic, here we refer to an EEC study (INEA,1992)

2 The number of farms it is quite high. We have calculated the optimal sample size of farms with 8 ESU or more, in 1986. The sample was stratified (ESU, Provinces) selected on the basis of "deff" analysis. The size was 713 versus 2089 farms observed in the actual survey.

3 Some studies were promoted by INEA to give a statistical ground to the analysis of data (Scala,1986): a sample was selected from the actual FADN.

4 Data quality was considered: i) detection and correction of outliers have been carried out according to both univariate and multivariate analysis; ii) farmers and interviewers were interviewed to check data collection procedures. Few outliers were found and data collection were found quite satisfying.

Kruskal and Mosteller (1979), in a comprehensive survey, review many definitions of representativeness, but we feel they do not give room to analytic solutions and can be considered only a philosophical justification.The second approach seemed a more general solution and has been explored in this work.

Randomization is the real basis for statistical inference as it eliminates personal choices and hence the possibility of a subjective selection bias. However, randomization is not such a general and simple concept, we must distinguish between experimental and observational frameworks (Smith 1983; Smith, Sugden 1988). Wold (1967) argued that experimental knowledge is reproducible knowledge arising from the control exerted by the scientist over the assignment of treatments. On the contrary, considering surveys, randomization means control over the selection of units. In any case, reproducibility acquires a very different meaning in finite population connected to social and economic phenomena. Moreover, in a social survey non-response and missing values, which introduce non-random selection, are quite large, so the basis for statistical inference appears weak even if a random sample is adopted. Finally, it must be considered that in many empirical works units from a social survey are necessarily selected without any randomization at all. For these reasons, designing and analysing in the social field is a very difficult task.

Starting from this point of view it is natural to accept the results obtained by Smith (1983), here summarised. "Non-random selection method can be ignored in a model based approach to inference if certain conditions are satisfied. The main condition that guarantees ignorability...is...(when) the selection of units does not depend on the measurement variables ... but only on the prior variables". In the light of this statement, the model based approach seems a possible way to face the problem. But Smith goes further and clarifies another possibility when the sample concerns observational studies and the units are chosen for convenience. "...The key assumption in generalising results from purposively selected observational studies to a wider population ...amounts to a statement that post-stratifying variables contain all the inferential information available in the design variables". From this quotation we draw that "For many non-random samples post-stratification is employed as a method for making descriptive inferences"

Furthermore, in a previous work Holt and Smith (1979) claim that "...(post stratification) is a device for protecting statistician's inference against those occasions when his randomization gives an unbalanced or unrepresentative sample." and conclude that "...post stratification (is)... a robust technique, relatively free of assumptions, which can be applied in a wide variety of situations. This conforms with our feeling that it is the structure of the population, rather than the sample design, which an estimator should reflect."

If we turn to the case of FADN, we find at least three main aspects that justify the possibility to refer to the above mentioned framework.

First of all, we have to stress that in agricultural surveys participation represents a quite general problem at least in Italy. It is quite meaningful that in the most important Italian sample survey on agriculture, the Structure of Farms, in front of a large sample rate and independent repetitions of the survey, the 38% of the farms observed have 50 Ha and more. Besides, in this group the sampling rate is 47%. It is fairly evident that, in spite of the independence over time, the same farms are included in different surveys. Moreover it should be considered that most Italian farms is small, and such farms cannot be considered proper farms, because economic activity and family management cannot be separated. Hence, farm size appears as an important *a priori* knowledge and we believe that even if the best sample were designed it would be very difficult to avoid some sort of non-random selection of sample farms. According to these considerations and taking into account the very few small farms included in FADN we decided to concentrate on farms having no less than 8 ESU (about 12,000 $). From now on we shall referto this subset.

The second point is the particular condition of agricolture production. Especially in Emilia Romagna the technology is fairly constant, the main factors of variability are connected with the quality of the land and its altitude, the "degree of specialisation of the area", on the one hand, and with types of farming, on the other. To check this aspect the cross-sectional patterns of the main variables of interest (production, value added and intermediate costs) have been observed. The analysis confirmed our assumptions.

The third point is that the above mentioned variables are the real factors driving the inclusion of farms in FADN. To check this aspect a twofold strategy was followed. First, a comparison among the sample distributions of some variables and the corresponding distributions from the survey on the Structure of Farms and population distributions (Agricultural Census) has been carried out. The variables considered were: altitude, Provinces; types of farming (field crops, fruits and vineyards, livestock, etc.), farms size: ESU. Second, a probit analysis was performed to evaluate how much the above variables explain the probabilty of inclusion of farms in the survey. We do not have room to present the results, in any case they confirm the importance of the variables considered in driving the farms inclusion in the smple, exluding altitude only. This means that groups of farms, characterized by combinations of the above three variables, are more prone to be included in the FADN.

The consequence we draw is that post stratification allow us to make inferences on population so that non-random selection can be ignored.

Post stratifying according to the above specified variables or some combination of them has to be evaluated in order to choose the most suitable criterion. Because the type of farming create many post strata with very few units, the post stratification has been performed according to Province and ESU.

The results obtained are in table 2 and show clearly the important effect of post stratification. Moreover, we would like to point out that post stratification estimates were submitted to experts of agricultural economics who considered the results very close to their expectations. About Utilized Area the difference of our estimate and the true value is 1.5%.

TAB 2 - Post stratified estimates of revenue costs and value added means and S E in 1986 (000 Lit)

	Revenue	*Cost*	*V.added*	*U.A.*
Mean				
Simple	125181	45306	79820	20.2
Post stratified	98837	35686	63251	17.3
Relative error				
Simple	2.52%	3.07%	2.50%	2.05%
Post stratified	1.76%	2.43%	1.86%	1.73%

4. Improving change estimates using repeated observations at successive time

We have already stressed the interest in repeated surveys and the contribution they give to the analysis of change. We have also shown (tab 1) that FADN is characterized by successive observations over time of subsets of farms. In this section our attention will be on the estimation of longitudinal parameters evaluating the opportunity to adop a Generalized Least Square (GLS) procedure restricted to take possible internal inconsistency (Fuller, 1990) into account.

Repeated surveys allow for revision of the estimates, according to new information collected each point time on the same units, as a consequence of the correlation they induce between the estimates over time (Wolter, 1979). As Patterson (1950) stressed, repetitions of a survey over time allow for efficient estimates of change which are obtained by efficient estimates on both occasions. But efficient estimates of the first occasion come from a revision on the basis of survey repetitions.

However, when some form of partial overlap exists, there are several alternatives to manage the new information for revision, depending on assumptions on population and on sample pattern repetition, Wolter (1979) and Fuller (1990) are good references.

If we have repetition and revision of the first occasions we must consider various subsets of units defined according to the patterns of permanence. The problem to solve is to combine over time the simple estimates arising from the various patterns of permanence to define composite estimates. The usual approach is to assume a linear relationship between simple and composite estimates. In any case, there are some alternative estimators of longitudinal parameters of the relationship to take into account constraints connected to the variables of interest (internal consistency) and characteristics of sample selection.

O.L.S. was considered, among others by Patterson (1950) and Jones (1980). This approach performs well in many situations, but OLS is no longer internally consistent when different subsets of observations are used for estimations. To face this problem, a different approach has been proposed (Wolter, 1979; Fuller, 1990). GLS, originally proposed by Jessen (1942), allows management of restricted estimates of parameters and ensures internal consistency.

Besides, as Fuller (1990) points out, by using an appropriate formulation of constraints, it is also possible to take into account the assumption made on the relationship between simple and composite estimates. Post stratification, which has a crucial role in this work, conditions the form of this relation.

In principle restricted GLS seems the most general and flexible approach, nevertheless, this approach could be very complicated to implement since the dimension of the model and the system of constraints increase very quickly as the complexity of the survey and the length of the time period increase. In this work, we have applied restricted GLS and have also checked if the difficulties arising in the practical implementation of restricted GLS are balanced by a real improvement of estimates.

In the following, we briefly introduce the GLS approach, the preliminary results of its application and some comparisons with other approaches are presented

The linear model is

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$$

where: $\mathbf{Y}$, vector of the simple estimates of the variable of interest; $\beta^* = (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1}\mathbf{X}' \mathbf{V}^{-1} \mathbf{Y}$, vector of composite estimates of the variable of interest; $\mathbf{e}$, error component, with $E\{\mathbf{e}\} = 0$ and $:E\{\mathbf{ee}'\} = \mathbf{V}$, the covariance matrix of the vector of simple estimators, assumed to be non singular; $\mathbf{X}$ is a matrix of dummies defined to take into account which simple estimates contribute to the estimate of each element of β.

The GLS estimator β restricted to be a linear combination of $\mathbf{Y}$, as described above is developed by Fuller:

$$\begin{pmatrix} \mathbf{X}'\mathbf{V}^{-1}\mathbf{X} & \Gamma' \\ \Gamma & \mathbf{0} \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \lambda \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{V}^{-1}\mathbf{Y} \\ \mathbf{G} \end{pmatrix} \mathbf{Y}$$

where Γ_i is a fixed row vector containing coefficients, of the constrained linear combination of β 's, $\lambda' = (\lambda_1, \lambda_2, ..., \lambda_b)$ are Lagrange's multipliers, **G** contains the coefficients of the linear combination (**GY**). The above equation defines the β's restricted estimator; the variance of the estimator is the upper k x k portion of

$$\begin{pmatrix} X'V^{-1}X & \Gamma' \\ \Gamma & 0 \end{pmatrix}^{-1} \begin{pmatrix} X'V^{-1} \\ G \end{pmatrix} V \left[\begin{pmatrix} X'V^{-1}X & \Gamma' \\ \Gamma & 0 \end{pmatrix}^{-1} \begin{pmatrix} X'V^{-1} \\ G \end{pmatrix} \right]$$

5. The implementation of GLS for FADN

Owing to the increasing amount of calculations involved and the emphasis on the experimental character of this step of our study, an empirical application will be carried out referring to a simplified frame, where the post stratification is limited to the size of farms, and the period of the permanence of the farms considered (period of overlapping) is limited to 2 successive waves of the survey (in practice, 1988-89). These two limitations require a comment. Post stratifying according to only one variable is not an optimal solution, as we have already shown. For the revision of estimates the period considered do not usually exceed four (Cicchitelli, Herzel, Montanari, 1992). In our case we have observed that the coefficients of adjustment decrease very quickly: on the third occasion the coefficients are very close to zero. This result was an important support to our decision to limit the application to 2 occasions and important as well for the practical utilization of FADN according to the procedure here proposed

The variables considered for estimation are value added (V), revenue (R) and intermediate costs (C).

In our case

$\beta = ({}_8R, {}_8V, {}_8C, {}_9R, {}_9V, {}_9C)$

where the elements of the row vector are composite estimates, indices refer to years 1988 and 1989.

To simplify the exposition, in the remaining part of this section the notation refers to only one stratum.

$\mathbf{Y} = ({}_8R_{8.}, {}_8R_{89}, {}_8V_{8.}, {}_8V_{89}, {}_8C_{8.}, {}_8C_{89}, {}_9R_{9.}, {}_9R_{89}, {}_9V_{9.}, {}_9V_{89}, {}_9C_{9.}, {}_9C_{89})$

the meaning of subscript in the vector is

8. = farms included only in 1988,

9. = " " " 1989,

89= " " in 1988 and 1989.

It should be to point out that the components of **Y** are internally consistent.

V is the covariance matrix of simple estimates obtained as in Battese, Hasabelnaby and Fuller (1989). Considering the non-randomness of our data could be preferable a model-based approach to obtain the estimation of the co-variance matrix.

X explains how the simple estimates are combined

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

Γ (1 on the principal diagonal and 0 elsewhere) expresses the constraints we are imposing on each β to ensure that it is a particular combination of **Y** determined on the basis of the post stratification.

G contains actual coefficients used for the post stratification.

The results of the application of the above procedure (restricted GLS) are presented in table 3, where the estimates for the averages of the variables of interest are compared with the corresponding estimates obtained by applying alternative approaches.

"Simple" (A), "Post stratified" (B), "Patterson" (C) - taking into account post stratification-,"Full GLS" (D) -no restrictions on estimates are imposed-,"Restricted GLS" (E) introduces constraints to ensure internal consistency and to take into account post stratification.

TAB. 3- Estimation of means of revenue (R), costs (C), value added (V) in 1989 with different techniques (000 LIT.)

	(A)	(B)	(C)	(D)	(E)
Mean					
R '88	121804	108438	107928	64212	108304
C '88	41532	37239	37342	19835	37184
V '88	80273	71198	70943	44405	71120
R '89	125294	109589	111194	67170	109402
C '89	45400	39669	39516	22127	39597
V'89	79895	69920	71058	45115	69805
Relative errors					
R '88	2.7%	2.2%	2.1%	1.7%	2.1%
C '88	3.6%	3.1%	2.9%	2.0%	3.2%
V '88	2.7%	2.3%	2.2%	1.8%	2.3%
R '89	2.5%	2.1%	1.9%	1.8%	2.1%
C '89	3.1%	2.7%	2.6%	2.1%	2.7%
V '89	2.5%	2.1%	2.0%	2.0%	2.1%

6. Conclusion

In this study we have investigated the possibility of exploiting an important source of data for analysis of agriculture and policy making. Three main

considerations led us to start this research. First, surveying agriculture is a quite difficult task, since there is a high rate of nonresponse and missing values are wide, hence the basis for inference are weak even if proper sample design is performed. Second, as a consequence of the first statement and considering the present lack of resources, it seems necessary to exploit as much as possible all available data sources. Third, a large part of farms included in the FADN is observed over time so, FADN can be considered a useful source for aggregate estimates and for estimates of individual change as well. In Italy, it is not possible to obtain this kind of data from other surveys.

According to a well known piece of literature, we have proposed post stratification to authorize inferences on the population of farms, finding very satisfactory results for the main variables of interest.

Special attention was paid to the longitudinal structure of the survey. A general frame to deal with this aspect, proposed for repeated surveys, was recalled and an empirical investigation was carried out to check the effectiveness of different approaches in our case. The results suggest some general considerations

Post stratification plays a very important role, in our case providing large gains in terms of efficiency of level estimates and, we can suppose, in terms of unbiasedness. Furthermore, its role is crucial when we consider longitudinal estimators. In fact, using non restricted GLS, where post stratification is not considered, the estimates of means cannot be accepted. On the contrary, techniques like restricted GLS, which take into account constraints imposed by post stratification and by the request of internal consistency, give much more satisfying results. Moreover the specification of constraints does not greatly increase the relative error of the estimates.

However, internal consistency, looking at our estimates, is respected even if constraints are not imposed. Generally, all the techniques used produce consistent estimates. Of course, this is not a general conclusion but only empirical evidence connected to some special case. This result suggests that when the problem of consistency can be relaxed, it seems not worth to using restricted GLS because, as we showed, they are more complicated to implement and the presence of constraints do not improve the efficiency. In such a case, and if high and stable individual correlations are found, as we checked in FADN, Patterson's is a better and more simple approach.

Nevertheless, the general advantage to use Restricted GLS seems evident when suitable information is not available about the problem of inconsistency or other assumptions requested by estimation techniques (as for Patterson). In such a situation restricted GLS is a very general and flexible approach with acceptable relative errors, and not too difficult to be implemented in practical situation as FADN.

References

Battese G.E., Hasabelnaby N.A., Fuller W.A. (1989), *Estimation of livestock inventories using several area and multiple frame estimators*, "Survey Methodology", vol 15, 13-27.

Cicchitelli G., Herzel A., Montanari G.E. (1992), *Il campionamento statistico*, Bologna, il Mulino.

Duncan G.I., Kalton G. (1987); *Issues of design and analysis of surveys across time*, "International Statistical Review", vol. 55, 97-117.

Dubgaard A., Grassmugg B., Munk K.J. (1984), *Agricultural Data and Economis Analysis*, "European Inst. of Public Administration", Maastricht.

Fuller W. A. (1990), *Analysis of repeated surveys*, "Survey Methodology", vol 16, 167-180.

Holt D., Smith T.M.F.(1979), *Post stratification*, "J. of the Royal Statistical Society "A, vol 142, 33-46.

INEA (1992), *Strategie famigliari, pluriattività e politiche agrarie, Bologna*, il Mulino.

Jessen R.J. (1942), *Statistical investigation of a sample survey for obtaining farm facts*, "Iowa Agric. Experiment Station Research Bulletin, 304.

Jones R.G. (1980), *Best linear unbiased estimators for repeated surveys*, "Journal of the Royal Statistical Society " B, vol 42, 221-226.

Kasprzyck D., Duncan G.I., Kalton G., Singh M.P.. (1989); *Panel surveys*, NY, J. Wiley.

Kruskal W., Mosteller F.(1979), *Representative sampling I,II,III,IV*, "International Statistical Review", vol. 47.

Patterson H. D. (1950), *Sampling on successive occasion wih partial replacement of units*, "Journal of Royal Statistical Society", B, 12, 241-255.

Scala C. (1986), *Il programma Inea per la progettazione e le analisi statistiche di un campione rappresentativo*, Roma, INEA.

Smith T.M.F.(1983), *On the validity of inferences from non-random sample*, "Journal of the Royal Statistical Society "A, vol 146, 394-403.

Smith T.M.F., Sugden R.A. (1988), *Sampling and assignment mechanism in experiments, surveys and observational studies*, "International Statistical Review", vol. 56, 165-180.

Statistics Canada (1992), *Design and analysis of longitudinal surveys*, Ottawa, November.

Wold H.O. (1967), *Non-experimental statistical analysis from the general point of view of the scientific method*, "Bull. of the International Statistical Institute", 36th Sess., book 1.

Wolter K.M. (1979), *Composite estimation in finite populations*, "Journal of the American Statistical Association", vol 74, pp. 604-613.