CHANGING INDUSTRY CODE SYSTEMS: THE IMPACT ON THE STATISTICS OF INCOME PARTNERSHIP STUDIES

Paul B. McMahon, Internal Revenue Service, P.O. Box 2608, Washington, DC 20013
paul.mcmahon@soi.irs.gov

ABSTRACT

The Statistics of Income Partnership Studies sampling plan depends on industry as one of four major stratifiers. This change caused the current edition of that design to use NAICS Codes that are thought to be successors to the specific SIC codes around which the design was fashioned. However, this is only an interim solution, for we do not have the conditional distributions of incomes, assets, or receipts under this revised industry coding scheme. Our focus, then, is on the design modifications, those in development, and the migration between the two coding schemes.

Key Words: Sampling Frame, Administrative Records

1. INTRODUCTION

The Statistics of Income series of studies are primarily designed for use by the United States Congress and Treasury Department in evaluating tax law provisions, both current and proposed. The partnership area studies have been conducted annually for almost 50 years now, with increased prominence during periods when tax shelters were popular. Over that time, a number of different sample designs have been employed, with the current strata outline dating back about a decade.

This design is now under review for four reasons. First, it is prudent to re-examine a sampling plan periodically, and this study's is past due. Second, the population is no longer contracting, as it was when the current design was implemented, and the length of the distribution's tail has also grown. Third, regulatory changes have forced us to make ad hoc changes to the existing design, and the impact is not well studied. One regulatory change that seriously affects the existing design is the switch from industry codes based on the Standard Industrial Classification Manual to codes based on the North American Industry Classification system.. This shift also changes the analysis, which is the fourth reason.

But first, some background on the administrative environment and the nature of the population. We will follow the path from the firm's classification through the processing environment and onto the sampling frame, with comments about the impact of the industry coding change. We will then examine the effect on the current design and the revision being planned.

2. BACKGROUND

The Statistics of Income studies use the tax forms filed with the Internal Revenue Service as questionnaires on economic issues. The use of administrative records for such purposes has many limitations, but the mandatory nature of the filing does constrain the nonresponse aspect. For partnerships, the filing of a return does not, under normal circumstances, arise from taxes due, but from the requirement that firms report the earnings of their owners, much the same as corporations report dividend income.

Partnerships are unincorporated businesses with more than one owner. Firms with this organizational structure have a form of business that falls between Individuals (acting as sole proprietorships) and highly structured Corporations. The definition for these firms is in the hands of the States, for as they decide what a corporation is; they also have a hand in determining what a partnership is. For example, there are a small number of businesses that have the interests in them traded on the open market like stocks, yet State laws call them Publicly Traded Partnerships. On the other hand, some joint operating agreements that appear on the surface to be partnerships, but under the laws are not. In addition, States forbid certain types of professional operations from becoming corporations, like law firms, forcing the confederations to become partnerships.

These examples might give the impression that the owners of the companies must all be individuals. Such is not the case. Corporations, tax-exempt organizations, individuals, and other partnerships may all be owners in any combination. The only constraint is that there must be two owners. Partnerships are not subject to an income tax directly; instead the income, tax liabilities, and credits are passed through to the owners.

These organizations are required to report their earnings and the distribution of that income and so on amongst the partners annually. The report includes a small space for a word, or perhaps two, describing the sort of business and product, and another box for the "Business code number." This business code was, before 1998, based on the Standard Industrial Classification (SIC) codes. The IRS's business codes were a consolidated subset of the SIC list. There were different tables of codes for different types of organizations, as well. Since Corporations, as an example, tended to have many manufacturing firms, there were far more codes for that division than were listed in the table for Partnerships, where that sort of business was scarce.

With the introduction of the North American Industry Classification System (NAICS), the Corporation and Partnership filers used the same list of industry codes. In part, this was because the Service was able to persuade the panel constructing the NAICS Codes that there were some sorts of businesses that had to have distinct classifications. Two such were for Regulated Investment Corporations (Open End Investment Funds, 525910) and Real Estate Investment Trusts (525930).

However, that list of codes was far too extensive for inclusion in the instruction booklet the filers are asked to follow, so the Service combined individual categories to fit their needs. The construction of the Service's codes followed the consequences of the tax and legal environment, rather than the economic concerns that influenced the development of the NAICS Codes. These legal concerns also led to a redefining of certain codes in the Insurance area as well.

3. ADMINISTRATIVE ENVIRONMENT

That list of codes is used by taxpayers in filling out their returns. Those records, whether on paper or, increasingly, electronic media, are then transmitted to the IRS. When the Service receives these records, it then converts the data to its own form of electronic record for tax administration purposes.

One might ask why not simply tabulate that data file. Certainly, with an entire population of less than two million firms, it would not be too great a burden on today's computers to produce tallies for any question at will. And, since the primary sponsors are permitted by law to view the individual records, a simple copy of the file would seem to suffice.

But the initial "Transaction Record" has faults, both in content and quality. First, beyond the Employer Identification Number, Tax Period, and industry, there are only 51 monetary amounts and 15 indicator fields available. We are ignoring the processing codes and other internal fields here, of course, as they have no bearing on our studies. Even so, not all of the germane fields are considered useable (because some are rarely applicable and, thus, overlooked), so we are left with 36 monetary and 5 indicator variables.

The quality issue also applies to the entity information at this stage of the process. Discrepancies between the Employer Identification Number and other information have not yet been identified, let alone researched and corrected. Among these bits of information untested at this point is the industry code.

The familiar problems endemic to self-reporting of codes are present, but there are other sources for apparent errors to arise as well. Table 1 shows the results for returns filed during 1999 and as of this writing for 2000. There is considerable error to these counts because some records get counted more than once (as they cycle through the error correction process). The total population processed through the sampling operation in 1999 was 1,972,000, meaning that the data are overstated by 1.6 percent.

Table 1: Partnership Returns Industry Coding

	1999		Through May 2000	
	Count	Percent	Count	Percent
Valid NAICS	1,619,000	80.8%	624,000	86.7%
Invalid NAICS	152,000	7.6	44,000	6.1
Unknown NAICS	98,000	4.9	24,000	3.3
Valid SIC	122,000	6.1	26,000	3.5
Invalid SIC	12,000	0.6	3,000	0.4
Unknown SIC	1,000	0.1	--	0.0
Total Returns	2,004,000	--	719,000	--

The categories in the table above are somewhat misleading. A "Valid NAICS" Code for this purpose was one that the IRS defined in its publications. "Unknown NAICS" would be records that did not have any industry code reported, but had a Tax Year of 1998 or later. Similarly, "Unknown SIC" means that no code was reported for records of Tax Years before 1998. The two "Invalid" lines contain a mix of those records that have SIC or NAICS codes that are not on the IRS's list, along with reporting and keying errors.

The trend clearly shows that the NAICS-based coding is improving. Last year only slightly more than 80 percent were valid (improving by only a half percent between May and year end), while, this year, the figure for May (2000) is closer to 87 percent. About half this increase arises from a diminishment in the number of SIC-based codes that appeared, but the rest arises from improved processing and reporting.

The reporting is a particular factor, because, while the filing deadline for nearly all partnerships is early in the year, it is quite straightforward to get an extension of 6 months, and not too difficult to get an additional 3 months beyond that. There are also some respondents who do not provide the industry information, and others who simply copy whatever they reported in that part of the form the previous year. Hence, even though the basic quality of the NAICS-based codes is encouraging, the processing had to accommodate both the NAICS- and SIC-based codes. The strategy IRS used was to insert two leading zeroes before the old SIC-based codes.

The administrative processing has one more step that needs to be addressed: Posting. The transaction records that gave rise to the statistics in Table 1 are matched to the IRS's Business Master File records. This posting process affirms the entity information in the transaction, updates the Master File, and allows certain permanent information from the Master File to be appended to the transaction record. The piece that interests us is that the last reported SIC-based code, posted prior to January 1999, is among the additional data.

This file from the posting process, with the enhanced transaction records, forms our sample frame.

4. CURRENT DESIGN

This sampling frame, then, has both SIC-based and NAICS-based industry codes, along with various monetary variables. But just as the industry coding is not quite what it seems, so it is with the other fields. Total Assets, for example, need not be reported for firms below a certain size if they meet some other, fairly relaxed, conditions. A similar situation exists for business receipts and net income. Here, in response to tax law definitions, the various sources of revenue are labeled "active" or "passive," then held to different procedures.

We will not go into the specifics of the current monetary classes, nor the statistical properties of this design, as that information was published elsewhere. The outline of the strata and the improvements in the coefficients of variation were described in McMahon (1995). The impact of the data abstraction process's quality was explored in McMahon (1996), and the use of permanent random numbers in sample selection in McMahon (1998).

In general outline, though, the immediate predecessor had four main sections. The first section contained a pair of strata for firms with assets of at least $100 million, or income or receipts of at least $25 million. The rest of the population was divided into the three remaining sections based on Industry. The set of strata reserved for Real Estate Operators (except developers) and Lessors of Buildings (SIC Code 6511) has been a staple of the Partnership design since the mid-Seventies. This predecessor design split the remaining population into a third section for Mining, Construction, Manufacturing, and Transportation companies (SIC Codes 1000 through 4999), and the fourth for Agriculture, Trade, Finance, Services, and companies without industry data.

The strata within these design sections depend upon Total Assets, Income, and Receipts. This classification has been in use for most of the decade, with only minor adjustments. We recently added an additional pair of strata at the upper end of the Asset and Income distributions to control the growth of the certainty class, for example. We have also had to set aside special classes for records identified as Publicly Traded Partnerships or for having been filed electronically.

But the key change was the conversion to the new six-digit industry code.

The instructions for abstracting the new industry code made provision for records reporting from previous years' industry classification. When one of the SIC-based codes was encountered during processing in 1999 (Tax Year 1998, mainly), two lead zeroes were inserted to distinguish them from the NAICS-based IRS industry codes. We used this industry code for the stratification, even though there was the old SIC code for continuous businesses.

The decision on how to handle the migration from one coding scheme to the other had to be made in the Fall of 1997, long before any data on the actual pattern of filing amongst the industries could be known. We also did not want to confound any analysis by instituting design changes beyond a minimum. Hence, we redefined the old categories in terms of the new-NAICS based codes, as shown in Table 2, below.

Table 2: Industry Groups Used in the Tax Year 1998 Sample Design

	Principal Business Activity Codes	
	Standard Industrial	North American Industry
Industry/Division	Classification	Classification System
Real Estate Operators	6511	531110 and 531120
Mining, Construction, Manufacturing, and Transportation	1000 through 4999	200000 through 350000, and 480000 through 519999
Farms, Trades, Finance, and Services	All other codes	

We also prepared for the advent of a large influx of records that had no industry information present. In this case, we deferred to the economists involved in the project. The parsing of the economic "receipts" into active and passive sources, while normally unhelpful, came to the fore. Where an amount of rent was present and the industry was missing or clearly wrong (in the ranges of less than 000100, 009000 through 110999, or greater than 820000), we declared those records to be in the Real Estate Operators strata.

Table 3: Partnerships by SIC Industry Division

	Estimated	Tax Year
Industry Division (SIC)	Tax Year 1997	1998
	Population	Population
Real Estate Operators	592,000	581,900
Finance	382,300	364,500
Services	311,000	320,100
Trade	173,000	152,400
Agriculture	127,100	118,600
Construction	72,100	67,300
Manufacturing	40,000	32,900
Transportation	30,900	25,500
Mining	28,000	23,700
Unknown		287,000

But why use the industry in the sample design at all? Since Real Estate Operators comprise about a third of the entire population (see Table 3, above), about 12,000 of the target 35,000 sample selections would have been in this one industry (about 2 percent of the estimates). By creating separate strata for them, we maintain the accuracy of that industry's estimates while halving the sample in that domain. We then use the roughly 6,000 records to reinforce estimates elsewhere. It is clear that Divisions like Transportation and Manufacturing are overshadowed and, thus, needed additional sample (beyond proportional allocation) to permit the level of analysis desired.

The data in Table 3 show another factor as well as the industry domination. The data for 1998 are derived from the historical SIC-based industry from the Master File, yet there is a segment of the population without this information. There can be only one reason: these are the records of new businesses and, thus, do not have prior year data. Thus, we cannot rely on the historical industry code for stratification.

5. DESIGNING WITH NAICS

The goals, in so far as the industry perspective is concerned, are first to identify any NAICS industry that would, if proportionately sampled, be allocated more sample units than are needed for analysis. Second, identify which, if any, are at risk of receiving too few observations. The third goal is to minimize the effect of records with unknown industry classification on the sample design.

Table 4 presents the migration observed in the frame using the records subjected to sampling during 1999 for the Tax Year 1998 Study. Since this is the first year of the new coding system, it is unsurprising that there would be a number of firms that received erroneous classification. This is, we believe, one source of the smallest frequencies reported below.

Table 4: Sampling Frame Transition From SIC to NAICS

		SIC Division									
NAICS Division	Total	Unknown	Agri-culture	Mining	Con-struction	Manu-facturing	Trans-portation	Trade	Finance	Real Estate	Services
Total	1,973,800	287,000	118,600	23,700	67,300	32,900	25,500	152,400	364,500	581,900	320,100
Raw Materials	134,300	9,300	91,200	19,200	500	2,600	1,300	1,300	4,400	1,600	2,800
Goods Production	129,500	22,500	1,700	200	51,400	17,100	400	3,900	23,000	3,200	6,000
Distribution	140,200	25,300	1,100	100	600	3,100	11,100	84,800	900	600	12,500
Information	17,800	3,600	*	*	100	1,700	3,800	600	300	200	7,500
Finance, etc.	936,600	100,500	5,100	400	2,500	400	1,700	3,800	275,800	507,900	38,400
Prof. Services	133,600	22,400	4,100	100	800	1,100	1,600	2,600	3,800	2,000	95,200
Education, etc.	36,700	5,000	*	*	*	*	100	200	300	300	30,800
Leisure, etc.	77,300	11,300	400	*	100	700	200	29,000	1,100	1,000	33,500
Other Services	56,500	9,200	1,500	*	900	500	400	2,100	1,200	400	40,300
Unknown	311,200	77,900	13,400	3,600	10,300	5,700	4,800	24,200	53,700	64,600	53,200

(Note: Rounded in hundreds, with an asterisk in cells where the count was less than 50)

The "Total" column shows where the largest industries are. The Finance Division, under the new system, has the lion's share of the population, with more than half arising out of the Real Estate Operators (SIC 6511) class. However, the non-Real Estate portion of Finance still contains about twice the population of the other divisions, so there may be other candidates as well. In fact, as Table 5 shows, there are four main components to the Finance Division.

Table 5: Largest Finance Division Industries

NAICS Industry	**Number of Firms**	**Assets (Millions)**
Other Financial Investment Activities	113,500	1,029,000
Lessors of Residential Buildings and Dwellings	285,300	459,000
Lessors of Non-Residential Buildings	237,000	596,000
Other Activities Related to Real Estate	116,700	253,000
All Other Finance Division	184,100	660,000

The two largest industries of Table 5 arise almost exclusively from the SIC Real Estate Operators category. This means that they share the same sorts of financial and business profile, and are compatible for stratification. The "Other Activities Related to Real Estate" firms are less closely tied to that profile, with only about 40 percent of their population having been previously identified as Real Estate Operators. Still, since they do share the same industry sector, there are reasonable gains to be had by including them in the special strata.

This does not appear to hold for "Other Financial Investment Activities." Less than 4 percent of that population were formerly identified as Real Estate and, as the Asset column shows, they clearly have a distinct distribution in that regard. Moreover, this collection of investment groups tends to have significant reported amounts of short and long term capital gains. These are areas with a history of large variability across the years, so it seems wise to leave them to a proportional representation, yielding about 1,600 sample observations.

On the other side of the coin, a proportionate share of the sample would lead to only about 300 records for the 4 published Information Division industries and a bit over 600 records for the Education, Health and Social Assistance

Division's 10 published categories. Such small sample sizes for the individual industries would support only the most cursory analysis. However, combining these NAICS Sectors means generating strata that are not homogeneous. Of course, the solution for this is clear: Post-stratification. In this case, the bias would be ignorable (effectively zero) because the post-stratification population data are collected during the sampling process.

But just as clearly, there is a problem in the category "Unknown." As we saw in Table 1, this population should be declining, but still we will need to make provision for this sizable group of records in the design. The real problem is that the population of records with unknown industry classification (at the time of sample selection) may contain about the same proportion of Lessors of Buildings as the population with known industries. The ad hoc plan used in the current sampling program, which simply uses the presence of rental income, tends to identify non-real estate operators too often.

Table 6: Real Estate Rents as a Proportion of Total Receipts

	No Rents **Reported**	**Under 70** **Percent**	**Between** **70 and 80** **Percent**	**Between** **80 and 90** **Percent**	**At Least 90** **Percent**
Lessors of Buildings	15.6%	4.1%	2.0%	37.8%	40.5%
Other Real Estate	59.0%	4.6%	1.4%	14.6%	20.3%
Agriculture	83.8%	4.0%	1.0%	5.0%	6.3%
Other Valid NAICS	96.5%	1.3%	0.1%	0.9%	1.1%

(Percentage of firms within each industry with a given proportion)

We propose to substitute an "80-percent rule," where a firm will be categorized (for sampling purposes only) as a Lessor of Buildings if the proportion of its receipts that are real estate rents exceeds 80 percent of the total. As Table 6 shows, this would have correctly identified over 78 percent of the known Lessors and about 35 percent of the industry "Other Activities Related to Real Estate." Only the Agriculture Sector would have a minor negative impact, misidentifying about 7,000 farms as belonging to the Real Estate areas, and this is not a significant concern to our users. Those companies that fail the 80-percent rule would be placed in one of the strata for sectors Trade, Raw Materials, and so on.

6. CONCLUSION

The conversion of the sampling strata from the SIC-based codes to the NAICS scheme was not entirely successful. The initial quality of the reported codes has improved, but the basic incompatibility of the two systems means that a completely new stratification plan is needed for the Partnerships Studies.

The replacement sample design will have five main categories. The largest firms and those with peculiar conditions will still have strata set aside for them, and the three industry groupings will be retained, although with updated particulars. What was once Real Estate Operators will now hold their successor industries, but the sparse industry group is entirely redefined. At this time, the full details of the revision are not known, for analysis of the data sets has only just begun.

7. REFERENCES

McMahon, P. (1995), "Statistics of Income Partnership Studies: Evaluation of the Expanded Sampling Plan," *Proceedings of the Section on Survey Research Methods, American Statistical Association,* pp. 650 – 655.

McMahon, P. (1996), "Non-Sampling Errors in Data Abstraction From Administrative Records," *Proceedings of the Section on Survey Research Methods, American Statistical Association,* pp. 184 – 189.

McMahon, P. (1995), "Longitudinal Estimates and Permanent Random Numbers in Administrative Records Studies," *Proceedings of the Section on Survey Research Methods, American Statistical Association,* pp. 709 – 714.

ON RECLASSIFYING INDUSTRIES FROM THE STANDARD INDUSTRIAL CLASSIFICATION SYSTEM TO THE NORTH AMERICAN INDUSTRY CLASSIFICATION SYSTEM[1]

Shawn D. Klimek and David R. Merrell, Center for Economic Studies, U.S. Bureau of the Census
Shawn D. Klimek, U.S. Bureau of the Census, 4700 Silver Hill Road, Stop 6300, Washington, DC 20233-6300
sklimek@ces.census.gov

ABSTRACT

The North American Industry Classification System (NAICS) presents significant challenges to users of Census Bureau economic data by limiting the time series dimension of the data. We develop an algorithm to reclassify the 1992 Retail and Wholesale Economic Censuses on a NAICS basis. First, we use a SIC-NAICS concordance to assign establishments in Standard Industry Classification (SIC) industries that match uniquely to a NAICS industry in 1997. Second, using establishment identifiers, we link establishments in operation in both 1992 and 1997. If the five-digit SIC codes match in the two censuses, we apply the 1997 NAICS code. The remaining establishments are classified in SIC industries that match to multiple NAICS industries. We construct the proportion of 1997 establishments migrating from an SIC to each NAICS code. Using these proportions as weights, the algorithm draws from the uniform distribution and randomly assigns the remaining establishments in a 1992 SIC industry to a 1997 NAICS industry.

Key Words: NAICS, SIC, Random Assignments

1. INTRODUCTION

The introduction of the North American Industrial Classification System (NAICS) was intended to classify industries more accurately by focusing on the processes of production rather than the products themselves. The idea is that the emergence of new technologies, new service industries, and new products posed significant challenges to the proper treatment of industrial classification. For a detailed history of the events surrounding the creation and introduction of the NAICS taxonomy, see North American Industry Classification System, United States 1997. While accurate and relevant industry coding is a laudable goal, the introduction of NAICS also poses significant challenges to users of economic data by limiting the data's time series comparability, and hence there is a compelling interest to maintain the time series dimension of economic data. This paper describes efforts underway at the Census Bureau's Center for Economic Studies to rectify the problem of comparing newly collected economic data (published under the NAICS system) to historical economic data (published under the SIC system).

This paper proceeds as follows. Section 2 details the algorithm developed to re-classify establishments included in the 1992 census from SIC industries to NAICS industries and presents some descriptive statistics that illustrate some of the salient features of the algorithm. Section 3 presents tables with preliminary aggregate tabulations for numbers of establishments, employment, and revenue for NAICS industries in 1992 and compares them to 1997. Section 4 discusses some refinements and extensions. Section 5 concludes, and Section 6 presents our list of referenced materials.

2. THE ALGORITHM

In this section, we describe the algorithm used to convert the 1992 Economic Census (EC) from the SIC basis to the NAICS basis. The algorithm was designed to create *aggregate* tabulations for the NAICS sectors 42, 44-45, and 72: wholesale, retail, and accommodation and food service, respectively. These sectors were chosen because the annual and monthly survey programs need to benchmark their data over a longer period than just 1997 and forward. However, the algorithm may be used more generally for non-manufacturing sectors.

The keys to implementing NAICS are the SIC bridge code. Consider the following example. The four-digit 1992 SIC 542100 was divided into three five-digit 1997 SIC bridge codes, 542110, 542120, and 542130. Each of these five-digit bridge codes then mapped to NAICS codes 44521010, 44522000, and 45439032, respectively. We use this concordance and the 1997 EC to construct the empirical distribution of 1992 SIC codes to 1997 NAICS codes. In this example, 74.6% of the SIC 542100 establishments are in NAICS 44521010, 23.1% in NAICS 44522000, and the remaining 2.3% in NAICS 45439032. We then integrate this distribution into the 1992 EC for the economic sectors mentioned above.

[1] Note: This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review by the Census Bureau than its official publications. This report is released to inform interested parties and to encourage discussion. Further, none of the tabulations for 1992 presented in this paper are official tabulations endorsed by the U.S. Bureau of the Census.

Table 1—Distribution of 1992 Wholesale and Retail SIC industries to 1997 NAICS industries

Number of 1997 NAICS codes matching to each 1992 SIC	Number of occurrences	Percent
1	218	78.98
2	47	17.03
3	10	3.62
4	1	0.004
Total	276	n/a

As Table 1 shows, almost 80% of 1992 SIC codes (six-digit level) are matched to a single NAICS code (eight-digit level). The first step of the algorithm assigns each establishment in these SIC codes the NAICS code in the concordance. We regard this method of assignment as the best, since it relies only on the industry classification of the establishment in 1992 and on the one-to-one industry correspondences in the concordance. Simply put, it requires the least a priori structure from us.

The second step in the algorithm links the remainder of the 1992 EC and 1997 EC at the establishment level using the Permanent Plant Number (PPN). The PPN is assigned to each establishment's physical location and remains unchanged, regardless of any changes in ownership or firm structure. Using the PPN, we identify establishments surviving from 1992 to 1997. If we observe the establishment in both 1992 and 1997 operating in the same five-digit SIC, then we assign the establishment in 1992 the same NAICS code it was assigned in 1997. We assume that consistent classification at the five-digit SIC level in both 1992 and 1997 implies that there are no changes in the type of industrial activity at the establishment level. Given this assumption, we consider this method a second best alternative to the one-to-one SIC-NAICS assignments described above.

The first two steps assign the majority of the 1992 establishments; however, a significant number of establishments still lack a NAICS code—even after step two. In the third (and final) step, we randomly assign the remaining establishments a NAICS code. When we merged the empirical distribution of SIC to NAICS codes, we included information about the proportion of establishments in the 1997 EC that are being classified from each SIC to a particular NAICS. For each establishment, the algorithm makes a uniform random draw and uses the proportions discussed above to weight each NAICS code with the appropriate mass of the distribution. Using this method we assign the remaining establishments a NAICS code. Table 2 below describes how the establishments in 1992 are assigned a NAICS code.

Table 2 —Number and Percent of Establishments in 1992, by Method of Assignment

Method of Assignment	Number of Establishments in 1992	Percent of Establishments in 1992
One-to-One Matches	1,282,603	61.84%
PPN Matches	184,072	8.88%
Random Assignment	607,246	29.28%

Clearly, random assignment is the least reliable method of assigning a NAICS code to an establishment since it will provide different assignments each time the algorithm is executed. However, we think that this method nevertheless is reasonable; Table 3 shows why. Table 3 shows that most of the establishments in an SIC that require random assignment are cases where more than 90% of the establishments migrate to a single NAICS industry. From Table 2, we observe 58 SIC industries that do not match uniquely to a NAICS code—representing 607,247 establishments in those 58 SIC industries requiring random assignment. Recalling the example at the beginning of the section, SIC 542100 and all of its establishments would be included in the "80% to 70%" row of Table 3, since 74.6% of the establishments are in NAICS 44521010 rather than the other two. Table 3 is supportive of the random assignment method since over two-thirds of the establishments in SIC industries are cases where the lion's share (over 90%) of establishments in 1997 are classified into a single NAICS industry. The main assumption underlying the random assignment portion of the algorithm is that of consistent industry composition. We assume that the correct probability of an establishment in 1992 is identical to the probability in 1997. There are at least two reasons why this assumption may not hold. First, rapid economic growth that differs across sectors and industries in the U.S. economy during the late 1990s could undercut this assumption. Second, the SIC industries most likely to be split up into several NAICS are those most likely to be experiencing rapid changes in industry structure. We feel that these two possibilities provide the basis for more work in providing even better aggregate tabulations for 1992 on a NAICS basis.

Table 3—Distribution of NAICS Industry Largest Shares for SIC Industries and Establishments

NAICS Largest Share	Number of SIC industries	Number of establishments
90% or more	5	410,962
90% to 80%	3	40,497
80% to 70%	3	64,202
70% or less	47	91,585
Total	58	607,246

3. RESULTS AND DISCUSSION

In this section, we present data (at the U.S. aggregate level) from three sources. The 1997 tabulations are from Census publications. The 1992 tabulations are constructed using the algorithm described in section 2. The SIC totals are from the Census Bureau's Advance Report.

We believe our numbers for the wholesale sector are quite reasonable. The growth from 1992 to 1997 in the number of wholesale establishments on an SIC basis was 4.59%. Comparing the numbers on a NAICS basis we compute 4.42% growth in the number of wholesale establishments. Second, employment on an SIC basis increases by 12.36%, where the numbers on a NAICS basis indicate employment increases by 11.08%.

The primary change implemented by NAICS was moving establishments "open to the general public" from the wholesale sector to the retail sector. This manifests clearly in the cross-sectional differences between SIC and NAICS in 1997. In 1997, there are 518,215 SIC wholesale establishments, but only 453,470 NAICS wholesale establishments. Differences between the establishments that remain in wholesale and those that moved to retail could explain the differences in the growth rates on an SIC and NAICS basis. Table 4 presents tabulations for the Wholesale Sector (NAICS 42) at the four-digit NAICS level.

Table 4—Four-digit Wholesale NAICS Industry Tabulations for 1992 and 1997

NAICS Industry	Establishments in 1992	Establishments in 1997	Employees in 1992	Employees in 1997	Revenue in 1992	Revenue in 1997
4211	30,942	29,328	343,749	375,761	368,575,847	553,352,124
4212	13,835	15,246	135,065	157,462	53,420,802	75,003,478
4213	12,772	14,267	137,554	155,535	69,792,059	89,175,875
4214	43,282	45,351	652,917	716,113	255,980,672	357,383,550
4215	11,248	12,583	138,042	174,029	118,321,902	150,493,610
4216	33,224	38,234	367,428	475,766	208,920,514	357,691,888
4217	19,517	21,194	190,776	219,233	63,869,994	92,189,762
4218	72,991	76,643	681,232	772,550	228,364,711	328,968,331
4219	34,737	37,783	306,163	351,839	140,387,696	185,455,758
4221	16,139	15,848	220,439	214,35 0	99,508,473	117,062,485
4222	6,069	8,053	157,855	190,127	129,306,287	203,147,771
4223	18,776	20,707	188,228	207,574	103,957,220	124,104,420
4224	42,622	41,760	805,929	854,919	499,946,049	588,970,062
4225	11,551	10,343	108,710	97,251	136,869,416	166,786,245
4226	14,193	15,920	147,010	165,768	132,471,184	128,923,496
4227	14,181	11,297	151,030	137,829	274,197,575	267,623,942
4228	5,259	4,850	141,821	151,677	59,487,322	69,703,203
4229	32,949	34,063	344,244	378,531	161,919,348	213,618,778
NAICS Totals	**434,287**	**453,470**	**5,218,192**	**5,796,557**	**3,105,297,071**	**4,059,657,778**
SIC Totals	**495,457**	**518,215**	**5,791,264**	**6,506,992**	**3,238,520,447**	**4,212,312,128**

For the retail sector, we make similar comparisons regarding the growth in the number of establishments and employment. There are two interesting things to note. First, growth in the number of retail establishments on an SIC basis was 2.61%, while we calculate 0.8% growth when comparing under the NAICS industry basis. Second, employment on an SIC basis increased by 15.98%, but on a NAICS basis we compute employment increased by only 14.78%.

In the 1997 cross-section, the 1,118,447 establishments in retail NAICS is dramatically lower than the 1,566,049 establishments in retail SIC. Given the transfer of establishments from wholesale to retail, this seems surprising. The large decline in the number of establishments in retail is primarily due to the creation of Sector 72, Accommodation and Food Service. However, in addition to NAICS Sector 72, some establishments move to other new NAICS service sectors and even to manufacturing. Table 5 shows our industry aggregate tabulations for the Retail Sector (NAICS 44-45) at the four-digit NAICS level.

Table 5—Four-digit Retail NAICS Industry Tabulations for 1992 and 1997

NAICS Industry	Establishments in 1992	Establishments in 1997	Employees in 1992	Employees in 1997	Revenue in 1992	Revenue in 1997
4411	43,052	49,237	922,932	1,138,995	349,832,714	553,652,292
4412	12,013	13,589	75,532	102,766	16,749,848	28,890,506
4413	57,231	59,807	413,518	477,200	54,532,471	62,824,978
4421	29,414	29,461	222,105	251,300	30,165,753	40,968,335
4422	32,667	35,264	187,244	231,545	22,278,040	30,722,478
4431	38,150	43,373	239,048	345,042	40,449,241	68,561,331
4441	73,190	71,916	767,805	952,296	138,659,107	195,888,196
4442	21,060	21,201	148,610	165,616	25,298,051	31,677,905
4451	107,404	96,542	2,525,407	2,643,608	332,215,630	368,250,471
4452	24,156	22,373	117,515	118,831	10,135,275	10,829,908
4453	31,386	29,613	132,989	130,635	20,319,081	22,684,120
4461	80,416	82,941	737,811	903,694	90,003,574	117,700,863
4471	128,369	126,889	817,263	922,062	154,043,396	198,165,786
4481	108,284	94,740	960,172	927,930	83,831,107	95,918,083
4482	37,206	31,399	184,415	185,803	17,883,367	20,543,252
4483	29,984	30,462	158,572	166,420	15,009,827	19,936,310
4511	46,929	46,315	319,956	362,973	31,456,522	41,415,227
4512	23,071	22,834	161,614	197,866	14,579,400	20,595,699
4521	10,346	10,366	1,585,742	1,795,577	168,370,441	220,108,157
4529	26,453	25,805	507,316	711,963	78,752,256	110,336,303
4531	27,341	26,200	122,114	125,195	5,719,237	6,555,088
4532	42,760	44,615	238,240	306,492	19,830,122	31573,035
4533	15,390	17,990	75,913	97,965	4,348,136	6,043,642
4539	32,460	41,033	142,677	223,334	16,630,362	33,937,396
4541	7,773	10,013	150,089	218,406	34,579,632	79,018,305
4542	6,391	7,070	69,628	66,348	6,330,079	6,884,497
4543	24,701	27,399	205,689	221,239	30,794,934	37,203,849
NAICS Totals	**1,117,597**	**1,118,447**	**12,189,916**	**13,991,103**	**1,812,797,603**	**2,460,886,012**
SIC Totals	**1,526,215**	**1,566,049**	**18,407,453**	**21,349,109**	**1,894,880,209**	**2,562,093,519**

The Accommodation and Food Service Sector (NAICS 72) is a new service sector created by NAICS; so, unlike the previous two sectors, we can make no comparisons of the SIC versus NAICS regimes. The sector is primarily composed of the two-digit retail major group SIC 58, Eating and Drinking Places, and a major group in services, SIC 70, Hotels, Rooming Houses, Camps, and Other Lodging Places. For the sector as a whole, we find that the number of establishments grew from 496,137 to 545,060, or 9.86%. The numbers were more dramatic for employment. Employment grew from 8,132,399 to 9,451,056, or 16.21%. Table 6 shows our industry aggregate tabulations for the Accommodation and Food Service sector at the four-digit NAICS level. Most industries show growth along all three dimensions with two exceptions. NAICS 7213, Rooming and Boarding Houses, shows small declines in both the number of establishments and employment. NAICS 7224, Drinking Places, shows a decline only in the number of establishments.

Table 6—Four-digit Accommodation and Food Service NAICS Industry Tabulations for 1992 and 1997

NAICS Industry	Establishments in 1992	Establishments in 1997	Employees in 1992	Employees in 1997	Revenue in 1992	Revenue in 1997
7211	41,736	47,079	1,456,093	1,645,666	67,200,771	94,965,838
7212	6,520	7,598	33,069	35,331	2,090,253	2,734,918
7213	3,561	3,484	17,530	15,597	720,302	754,105
7221	170,030	191,245	2,983,807	3,641,402	85,011,492	112,450,172
7222	191,481	214,767	2,908,000	3,326,543	85,823,575	107,780,513
7223	26,961	28,062	429,854	464,870	16,203,609	19,407,810
7224	55,848	52,852	304,046	321,294	11,113,777	12,292,709
NAICS Totals	**496,137**	**545,060**	**8,132,399**	**9,451,056**	**268,163,779**	**350,389,065**

4. REFINEMENTS AND EXTENSIONS

In this section, we propose three refinements to our algorithm to increase accuracy and provide better measures of measurement error due to random assignment.

We currently use only plant-level information to assign NAICS codes to establishments. This can create a problem illustrated by the following hypothetical example. In 1992, a multi-unit firm with 100 establishments operates in SIC 541140, Grocery Stores. In 1997, the same firm operates 110 establishments in NAICS 44511020. Under our current method of PPN matching this is not a problem if the 100 establishments from 1992 continue to operate in 1997. However, suppose that the firm closes 20 establishments (i.e., the PPN doesn't appear in the 1997 EC), and then opens 30 new establishments (i.e., 30 new PPNs appear in the 1997 EC). Under our current algorithm, we assign 80 establishments to 44511020, but we randomly assign NAICS codes to the 20 plants that close between 1992 and 1997. Currently, even when *all* of the establishments of the firm remain in the same five-digit SIC from 1992 to 1997 (and just one NAICS in 1997), it is possible that the algorithm assigns these exiting establishments to an inappropriate NAICS code in 1992—simply because they will fall into the class of establishments requiring random assignment. To correct this potential problem with the algorithm, we generate firm level data in 1992 and 1997. We restrict the sample only to firm records with more than one establishment appearing in only one SIC (five-digit) in 1992 and only one pair of SIC (five-digit)-NAICS (eight-digit) codes in 1997. Matching across the two years at the five-digit SIC level and keeping only extant firms will generate a dataset at the firm level with the appropriate firm level NAICS code. We add this step into the algorithm after the one-to-one and PPN matching steps, but before the final step of random assignment. The effect, we believe, will be to reduce the number of establishments that require random assignment.

We currently use only the percent of establishments in 1997 that move from a SIC to a NAICS code as the probability an establishment in 1992 moves to that same NAICS code. In order to use all of the information available, we plan to estimate the probability of a 1997 establishment moving from an SIC to a particular NAICS code—using traditional limited dependent variables techniques such as multinomial logistic regression. Using the parameter estimates from the model on 1997 data, we then generate a revised probability of being assigned to a particular NAICS code for each establishment based on its 1992 characteristics. This approach assumes that firm characteristics have the same effect on the probabilities across the two census years. We then make a random draw from the uniform distribution (weighted by the share of the probability mass estimated from the regression models) for each establishment and assign the appropriate NAICS code.

One additional weakness of the algorithm in section 2 is that each time the algorithm is executed an establishment can be assigned a different NAICS code depending on its random draw. This is true for all establishments assigned by the random assignment method. We propose implementing a bootstrapping method to simulate the true aggregate tabulations. To do this, we simply repeat the random assignment process (with the multinomial logit estimates mentioned above) a large number of times, each time generating the aggregate tabulations of interest. Our final estimate of the "true" tabulation is the mean of this distribution. In addition to the mean, we also will have estimates of the variance and other higher moments of the distribution. We expect this bootstrapping method to make very little improvement in assigning the *number of establishments* for NAICS industries in the 1992 EC. Given the assumption of the uniform distribution and identical weights on each NAICS code for each establishment across iterations, we believe that there should be low variance with respect to the number of establishments. However, depending on the heterogeneity of employment and revenue across establishments in 1992, the random assignment step could have potentially large effects on these aggregate tabulations. We expect our estimates of the variance and other higher moments to provide some insight on the severity of measurement error for these variables.

Current efforts focus entirely on the 1992-1997 reclassification in the wholesale, retail, and accommodation and food service sectors. We anticipate that future efforts will focus not only on the refinements discussed above but also on extending this work to other sectors such as manufacturing, services, and finance, insurance, and real estate. We think there is a compelling interest in extending the breadth of the NAICS conversion efforts to as many economic sectors as possible. Additionally, we anticipate that future efforts will aim toward maximizing the amount of historical data files that are converted. We feel that the algorithm developed to convert the 1992 Economic Census is sufficiently general to allow us to apply it to historical economic census data sets in many economic sectors.

5. CONCLUSION

In this paper, we present a methodology to reclassify industries from the Standard Industrial Classification system to the newly introduced North American Industry Classification System. This algorithm, designed solely for use in generating aggregate tabulations, uses three components to classify 1992 SIC codes to NAICS codes. First, cases were identified where SIC codes had unique correspondences to NAICS codes; in these cases, we merely assigned the unique correspondence backward to 1992. Second, for cases in which there are extant establishments between 1992 and 1997 and for which we observe those extant establishments producing in the same five-digit SIC in both 1997 as in 1992, then we assign the 1997 NAICS code to the 1992 establishment observation. The clear majority of industry assignments are made using these first two steps. Finally, for establishments that existed in 1992 but not in 1997 and for which there are no uniquely corresponding SIC to NAICS codes or establishments that switched industries, we assign a 1992 establishment observation a NAICS code based on random draw from a uniform probability distribution weighted by the proportion of 1997 establishments that migrated from the 1992 SIC to a given NAICS industry. Only about 29% of all establishments in NAICS 42, 44-45, and 72 are assigned to 1992 establishments observation using this method.

Recognizing that no reassignment algorithm will be perfect, we propose a number of refinements and extensions. These extensions range from modeling the probability that an establishment will migrate from one NAICS versus another based on establishment and firm observables (e.g. product lines and class of customer information) to generating standard errors associated with multiple iterations on the random assignment cases. Additionally, we think that we could refine our estimates by reducing the number of random assignment cases. This can be done by aggregating to the firm level (for establishments belonging to multi-unit firms and that are subject to random assignment) and imposing NAICS codes based on the firm's NAICS code. Finally, we propose to extend our work to include more historical non-manufacturing Economic Census years and to extend our efforts toward reclassifying manufacturing data as well.

6. REFERENCES

Executive Office of the President (1998), *North American Industry Classification System, United States 1997*, Washington, D.C.: U.S. Office of Management and Budget.

Bureau of the Census (1999), "Accommodation and Food Services—Geographic Area Series," *1997 Economic Census,* Washington, D.C.: U.S. Department of Commerce.

Bureau of the Census (1999), "Advance Report," *1997 Economic Census,* Washington, D.C.: U.S. Department of Commerce.

Bureau of the Census (1999), "Retail Trade—Geographic Area Series," *1997 Economic Census,* Washington, D.C.: U.S. Department of Commerce.

Bureau of the Census (1999), "Wholesale Trade—Geographic Area Series," *1997 Economic Census,* Washington, D.C.: U.S. Department of Commerce.

THE TRANSITION OF THE U.S. BUSINESS REGISTER TO NAICS

Michael E. Kornbau, Carl A. Konschnik, Paul S. Hanczaryk, U.S. Bureau of the Census[1]
Michael E. Kornbau, U.S. Bureau of the Census, Washington, DC 20233
michael.e.kornbau@ccmail.census.gov

ABSTRACT

The Bureau of the Census maintains a Business Register of all active business establishments with employees. In addition to Census Bureau internal sources, we use administrative record sources from three external government agencies for industry classification of businesses. These agencies are in a transition from supplying us with SIC-based codes to NAICS-based codes. The 1997 Economic Census provided full SIC and NAICS classifications for most existing employer establishments, but we rely on administrative records for classifying all new and some existing employer establishments. This paper presents how we update the Business Register with industry classifications, and describes the specific challenges we face in obtaining the best possible industry codes during the transition.

Key words: SSEL, administrative records, transition, SIC

1. INTRODUCTION

The Bureau of the Census maintains a Business Register, which we call the Standard Statistical Establishment List (SSEL), of all active business establishments with employees in the United States. The SSEL is used for statistical purposes only and acts as a frame for economic surveys and as a mailing list for economic censuses. In addition, the SSEL is used as the basis for statistical tabulations -- employment and payroll data are tabulated from the SSEL to produce the annual County Business Patterns statistics. Because data users are interested in the different types of industrial activity, a critical element of the Business Register is the industrial classification of each establishment.

The initial SSEL in 1974 used the 1972 Standard Industrial Classification (SIC) system for classifying establishments. The SIC system was updated in 1977 and 1987, and the SSEL incorporated these changes. The year 1997 saw the introduction of the North American Industry Classification System (NAICS), a system that introduced major changes in classification concepts and entirely new numerical classification codes. The SSEL is currently in a transition state between the old SIC system and the new NAICS system of classification.

This paper will discuss the external and internal sources for our industry classifications, how the sources are converting to NAICS, and how we're incorporating these changes into the SSEL. It will conclude with our plans for improving classifications on the SSEL over the next several years.

1. SOURCES OF INDUSTRY CODES FOR BUSINESS ESTABLISHMENTS

2.1. Social Security Administration (SSA) Business Birth Codes

A new business that is planning to hire employees is required to obtain an Employer Identification Number (EIN) from the Internal Revenue Service (IRS) in order to file employer-related tax returns. A business does this by completing IRS Form SS-4, Application for Employer Identification Number. The SS-4 is a one-page application that includes several questions pertaining to the type of business:

[1] This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review by the Census Bureau than its official publications. This report is released to inform interested parties and to encourage discussion.

- Principal activity
- Is the principal business activity manufacturing? (If "Yes", principal product and raw material used)
- To whom are most of the products or services sold?

After assigning an EIN, the IRS provides the SSA with copies of the Form SS-4 for geographic and industrial classification. The SSA has engaged in industry coding since 1936-37, and has shared classifications with the Census Bureau since 1948 (U.S. Department of Commerce and Social Security Administration, 1999).
Prior to 1999, the SSA assigned SIC codes through a manual coding process involving a staff of coders. Under an annual reimbursable agreement, the SSA provided the Census Bureau with SIC codes for about 100,000 new EINs each month. In 1998, the SSA and the Census Bureau produced a Memorandum of Understanding (MOU) for sustaining SS-4 coding within SSA using NAICS codes. The staff of SSA coders switched over to assigning NAICS codes with SS-4's received in 1999.

The new, automated NAICS coding system at SSA uses the Coding Assist and Ruling Request System (CARRS) developed by Statistics Canada. The CARRS allows a coder to enter a business description on a personal computer equipped with the software and receive a list of potential NAICS codes. The coder decides which code best classifies the business. The SSA coders were trained on the use of CARRS in January and February 1999 by the Census Bureau's classification staff. The use of CARRS for NAICS classification was a major change for the SSA coders. Many coders had years of experience with manual SIC coding, and were not accustomed to working with a personal computer.

The transition to NAICS at the SSA was very successful. Our first receipt of SSA NAICS codes was April 1, 1999. We continue to receive NAICS codes monthly. The unclassified rate, or the rate at which no code could be assigned based on the information present on the SS-4, dropped from 30 percent under the SIC coding system to less than 10 percent under NAICS. This precipitous drop in the unclassified rate is due in large part to the success of the CARRS system in assigning codes, but also to the increased commitment of SSA staff in assigning codes to each SS-4 return.

2.2. IRS Principal Business or Industry Activity Codes

Employer and nonemployer businesses and organizations are required to file an annual income tax return to the IRS. The possible tax returns include the following:

- Form 1040, Schedule C, Profit or Loss From Business (Sole Proprietorship)
- Form 1065, U.S. Partnership Return of Income
- Form 1120, U.S. Corporation Income Tax Return

These tax returns request a principal business activity (PBA) code, and a description of the principal business and the principal product or service. A list of possible PBA codes is provided to the tax filer in the tax instructions. The filer chooses the code that best describes the business from the list.

Title 13, United States Code, Section 6, authorizes the Census Bureau to acquire and use information from other Federal departments and agencies for statistical purposes. Under this provision and under Internal Revenue Code, we are able to receive business classifications in the form of PBA codes from the IRS.

For tax years 1997 and before, the PBA codes reflected the SIC system, but were sometimes, but generally not actual SIC codes. For example, the corporation PBA code '1510', representing general building contractors, covered businesses in SICs 1521, 1522, 1541 and 1542, while the PBA code '1531', for operative builders, covered businesses in SIC 1531. Starting with 1998 tax data, the PBA codes are either partial or full six-digit NAICS codes. The sole proprietorship (Form 1040, Schedule C) return allows for 301 codes, while the corporation (including S corporations) and partnership returns both allow for 422 codes. The lists for corporations and partnerships are identical.
The Bureau of the Census and the IRS have evaluated the accuracy of these self-assigned PBA codes under the SIC system (Konschnik, et al., 1993). We are currently reviewing the quality of the NAICS PBA codes.

2.3. Bureau of Labor Statistics (BLS) NAICS Codes

The Bureau of Labor Statistics (BLS) maintains a business list, called the Business Establishment List (BEL), which is similar to the SSEL. One of the major uses of the BEL is for the BLS' Covered Employment and Wages (ES-202) Program. The source of much of the BEL data is the Unemployment Insurance (UI) system of state employment security agencies, or SESAs. Thus, the data are obtained at a state level as compared to the national level for the SSEL.

A new employer is required to file a form with their state that includes a written description of their type of business. BLS assigns an industry classification, usually a complete code, based on this description. After the initial classification, BLS re-contacts about one-third of the UI-covered employers in an annual refile survey. Virtually all establishments are included in the survey within a three-year period. Industrial activity is reviewed among other items. The industry codes are crucial to the BLS. An SIC quality assurance program was developed in 1992 to ensure accurate industry code assignments (Farmer and Searson, 1995).

The Census Bureau and the BLS signed an MOU in 1991 for the BLS to provide SIC codes to the Census Bureau for a specified number of unclassified and partially classified establishments (U.S. Department of Commerce and U.S. Department of Labor, 1991). Under the agreement, the Census Bureau creates two files of EINs from the SSEL on a quarterly basis: one file with EINs which do not have an industry classification, and one file of EINs with a partial classification. The BLS returns industry classifications from the BEL for the EINs to the Census Bureau.

The BLS is converting the BEL to a NAICS basis through a three-year process, starting in November 1998 with EINs that have a 4-digit SIC code converting directly to a 6-digit NAICS code.

2.4. Economic Censuses and Surveys

The Bureau of the Census conducts an economic census of employer business establishments in years ending in ‘2' and ‘7'. In late 1997, we mailed around 3.5 million establishments an economic census form. Each form includes questions regarding industry classification. Based on responses to these questions, Census assigns an industry code.

For 1997 the assigned industry code took the form of a six-digit “bridge” code.
The bridge code is a four-digit SIC code plus two additional digits that make it possible to map to a unique six-digit NAICS code. Table 2.4.1 displays an example.

Table 2.4.1 SIC to NAICS Bridge Codes for SIC=5441

SIC	SIC Description	Bridge Code	NAICS	NAICS Description
5441	Candy, Nut and Confectionary Stores	544101	445292	Confectionary and Nut Stores (Retail)
5441	“	544102	311330	Confectionary manufacturing from purchased chocolate
5441	“	544103	311340	Nonchocolate Confectionary manufacturing

For each bridge code, there is a corresponding unique SIC-NAICS combination. This makes it possible to assign either a full 4-digit SIC code or a full 6-digit NAICS code to a particular establishment. Thus, it is possible to create estimates by SIC to compare with previous censuses and surveys, or to produce estimates reflecting the new NAICS structure.

Industry classification is a high priority for the Bureau of the Census. The assigned industry classifications are reviewed and edited. Imputation procedures, such as hot deck imputation. assign industry codes to establishments not responding to the economic census mailout or followup.

A certain population of employer business establishments do not receive economic census forms. These include establishments in out-of-scope industries such as agricultural services, post offices, and railroads. They also include many construction establishments, which are sampled for the economic census, and many very small establishments in the retail trade sector. We mailed a classification form to each of these establishments not included in the census mailout if we were uncertain of its classification. This mailout included 1.6 million classification forms in 1997.

Between census years, several economic surveys collect industry classification. The Quarterly Business Birth Survey (BSS) is conducted for several sectors of the economy. A mailout questionnaire collects industry information for a sample of new employer establishments. The goals of the BSS are to track company organization and industry classification. From this effort, we receive 7,000 to 8,000 classifications per quarter. The BSS is currently on a bridge code basis. Other large-scale Census Bureau surveys, such as the Company Organization Survey (COS), the Annual Survey of Manufactures (ASM) and the Current Industrial Reports (CIR) also update the SSEL with industry codes.

3. THE SSEL CONVERSION TO NAICS

The Business Register is currently in a transition stage where an establishment may have a mixture of SIC and NAICS codes from the sources covered in section 2. From this mixture of codes, our surveys require the best classification on an SIC **and** a NAICS basis. This leads us to use a bridge code as our primary code. Starting in 1998, the 4-digit SIC field on our Business Register was converted to a 6-digit bridge code, which maps to a unique SIC-NAICS combination.

3.1. SIC - NAICS Hierarchy

As we receive codes from several sources, it is necessary to decide upon one industry classification from the available sources. Sometimes these sources may disagree. Under the SIC system, we employed an SIC hierarchy, which selects an SIC code from available sources based on the source with the best reliability. The following was the SIC hierarchy:

1. SIC code reported in an economic census
2. SIC from a current survey
3. SIC code from a Classification Card (Census years only)
4. SSA SIC code
5. BLS SIC code
6. SIC code from name coding or description coding program
7. IRS PBA code

In the hierarchy, a code from a higher source (lower number) is used over a code from a lower source. For example, if an establishment has the following codes:

BLS SIC Code = 5812
SSA SIC Code = 6512
Economic Census SIC code = 7011,

the assigned SIC is 7011, since it comes from the highest source, an Economic Census code.

During the transition stage, we continue to assign our “best” SIC code from available SIC codes, using the SIC hierarchy. In addition, we assign a “best” NAICS code from available NAICS codes, using the same hierarchy scheme, except replacing ‘SIC’ with ‘NAICS’. If an establishment has a particular SIC-NAICS combination that corresponds to a unique bridge code, then we can assign a bridge code. However, in many cases either the SIC or the NAICS code is a partial code, or the SIC-NAICS combination is invalid.

3.2. Bridge Code Assignment

For the final 1998 version of the Business Register, we introduced a probabilistic assignment of a bridge code for any establishment with a partial classification.

The probabilistic assignment selects a unique bridge code from a list of possible bridge codes corresponding to the best SIC and the best NAICS code for the partially-classified establishment. There will only be one possible selection if both the SIC and the NAICS are fully classified, and they agree. Otherwise, we need to select a code using a probability distribution of all possible codes. We generate a random number between .000 and .999. We select a code by matching the random number against the cumulative probability distribution of possible bridge codes. A particular code is selected when the random number is less than the cumulative probability of the code but greater than or equal to the cumulative probability of the preceding code on the list of possible codes. The probability distribution is based on the distribution of establishments in the 1997 Economic Census with reported bridge codes, that were designated as a birth in the three years preceding the census.

As an example, suppose an establishment has an SIC code of '4522' and a partial NAICS code of '481200'. Table 3.2.1 shows a list of possible bridge codes that correspond to an SIC code of '4522' and a NAICS code of '481200'.

Table 3.2.1 Using a Distribution to Assign a Bridge Code

SIC Code	NAICS Code	Bridge Code	Estab. Count	Cumulative Probability
4522	481211	452201	84	0.183
4522	481212	452202	363	0.978
4522	481212	452209	10	1.000

We generate a random number for the establishment. Let's say the random number is 0.488. The number 0.488 is greater than 0.183, but less than 0.978, so we assign the bridge code of '452202' to the establishment. The random number generator uses the EIN as the seed. We require that the random number generator produce the same random number for an EIN in different program runs, so that we will not assign a different bridge code for every run.

Along with the selected bridge code, we also store on the SSEL the estimated probability that the code is truly the correct code. In the example above, the estimated probability would be 0.978 - 0.183, or 0.795.

For this assignment scheme, if the best SIC and the best NAICS disagree (were an invalid combination), then we would select either the SIC or the NAICS, depending on which has the higher source and is the most current, and select the bridge code solely from the one code.

Each month, we run the probabilistic assignment process for every partially classified, single-unit establishment on our Business Register. We also run the probabilistic assignment for any single-unit establishment that was previously assigned through this process, but now has an updated NAICS code. Unclassified establishments and multi-units are excluded. Most establishments were fully classified in the 1997 Economic Census, so the probabilistic assignment covers mostly new establishments, out-of-scope establishments and small establishments not selected for the census. The May 2000 Business Register had over 1.4 million active (reporting payroll in 1999) establishments with a bridge code coming through the probabilistic assignment. For 55 percent of these establishments, the available SIC and NAICS codes allow us to assign a full 6-digit bridge code directly, without more than one possible choice. The remaining 45 percent required an assignment from a list of more than one possible bridge code. For these 45 percent, the estimated probabilities that the assigned code is the correct code are as follows:

- 25% are assigned a bridge code with probability of 0.84 or higher
- 50% are assigned with probability of 0.52 or higher (median probability)
- 75% are assigned with probability of 0.18 or higher

The effect of the probabilistic assignment is that approximately 11 percent of current active establishments on the Business Register have an imputed bridge code, accounting for 2.5% of 1999 payroll.

4. THE FUTURE OF INDUSTRY CODES ON THE BUSINESS REGISTER

The Business Register is scheduled to go through a major redesign in 2002. By that time, the Register will be on a NAICS basis, and we may want to eliminate the use of any SIC codes in setting the NAICS code.

We have several goals for the 2002 redesign and beyond. One goal is to get quality measurements of our NAICS code sources. In the past, we've conducted evaluations of IRS PBA codes and BLS SIC codes. We will need to update these evaluations with IRS, BLS and SSA NAICS codes. The evaluations will compare codes from administrative records sources with codes we collect through the Quarterly Business Birth Survey or the economic census.

We also plan to change the hierarchy system of assigning NAICS codes. The current system always assigns a code from one source over a code from another source that is lower on the hierarchy. No consideration is given to the age of the code from the higher source. This may be a problem if we receive an SSA code assigned at business startup, but with the type of business changing over time. If we don't receive a code with a higher source, the code is not updated. We also haven't considered differences in quality by sector or industry. There is potential to add other sources for industry codes, such as internet resources. In these cases, we need a method to determine coding accuracy without spending years in evaluations before usage.

We are considering revising the method we use to assign a NAICS code, including the probabilistic assignment process. We can use additional characteristics about an establishment such as geographic location, business name and size in making a decision on a six-digit NAICS code

An additional goal is to introduce a quality control program for industry codes. This would involve evaluating codes as we receive them, providing feedback to our sources such as the SSA or the BLS, and a review of processing steps that post codes to the Register. We currently have several quality assurance programs to review incoming administrative record data. However, this is at the aggregate level, and does not check the quality of data from individual records.

4. REFERENCES

U.S. Department of Commerce, Bureau of the Census (1979), *The Standard Statistical Establishment List Program*, Technical Paper 44

U.S. Department of Commerce, Bureau of the Census and the Social Security Administration (1999), Memorandum of Understanding Between the Bureau of the Census (BOC) and the Social Security Administration (SSA) for Sustaining Employer (SS-4) Coding Within SSA, signed 12/23/98 by Peter Wheeler, Associate Commissioner, Office of Research, Evaluation, and Statistics at SSA, and signed 1/12/99 by Tom Mesenbourg, Assistant Director for Economic Programs at BOC.

Konschnik,C, J. Black., R. Moore, and P. Steel (1993), "An Evaluation of Taxpayer-Assigned Principal Business Activity (PBA) Codes on the 1987 Internal Revenue Service (IRS) Form 1040, Schedule C," *Proceedings of the International Conference on Establishment Surveys, American Statistical Association*, pp 745-750.

U.S. Department of Commerce, Bureau of the Census and the U.S. Department of Labor, Bureau of Labor Statistics (1991), Memorandum of Understanding Between the Bureau of the Census and the Bureau of Labor Statistics, signed 4/18/91 by Barbara Everitt Bryant, Director, Bureau of the Census, and signed 4/19/91 by Janet L. Norwood, Commissioner, Bureau of Labor Statistics.

Farmer, Tracy E., Michael A. Searson (1995), "Use of Administrative Records in the Bureau of Labor Statistics' Covered Employment and Wages (ES-202) Program," *Proceedings of the 1995 Bureau of the Census Annual Research Conference,* pp 198-235.

IMPLEMENTING THE NAICS FOR BUSINESS SURVEYS AT BLS

Gordon Mikkelson, Teresa L. Morisi, George Stamas, U.S. Bureau of Labor Statistics
George Stamas, Bureau of Labor Statistics, Suite 4985, 2 Massachusetts Ave NE, Washington, DC 20212
Stamas_g@bls.gov

ABSTRACT

To implement the North American Industry Classification System (NAICS), the Bureau of Labor Statistics and State partners are assigning NAICS codes to the approximately 8.2 million employers covered by State unemployment insurance (UI) laws. Employer UI reports are the basis of the Longitudinal Data Base (LDB), which serves as the frame for BLS establishment surveys. The NAICS conversion includes a multi-year process of gathering information from employers in order to assign NAICS codes. The collection procedure allows for interim assessment of the effect of the NAICS conversion on industry classification and BLS products. When employers do not provide adequate information for industry classification, BLS will assign NAICS codes based on the distribution of those codes across other establishments with the same Standard Industrial Classification (SIC) and other characteristics. These procedures will be applied to current and, to the extent feasible, historic data on the LDB including establishments that are out of business. This provides a frame for surveys requiring stratification by NAICS and aids in the conversion from SIC to NAICS for ongoing surveys. In addition, the availability of a continuous history with NAICS codes will permit seasonal adjustment and other time-series analysis of the data.

Key Words: Industry classification, Sampling frame, Nonresponse

* All opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics.

1. INTRODUCTION

The Covered Employment and Wages Program, commonly referred to as the ES-202 program, is a cooperative program between the Bureau of Labor Statistics (BLS) of the U.S. Department of Labor and the State Employment Security Agencies (SESAs). The ES-202 program produces a comprehensive tabulation of employment and wage information for workers covered by State unemployment insurance (UI) laws. Employer UI reports also are the basis of the Longitudinal Data Base (LDB), which serves as the sampling frame for BLS establishment surveys. For more information on the ES-202 program, see the *BLS Handbook of Methods,* Bulletin 2490 (Bureau of Labor Statistics, April 1997).

In order to implement the North American Industry Classification System (NAICS) for the ES-202 program, the BLS and its State partners are assigning NAICS industry codes to the approximately 8.2 million employers covered by UI laws. The NAICS conversion includes a multi-year process of gathering information from employers in order to assign NAICS codes. Whenever employers do not provide adequate information for industry classification, BLS assigns NAICS codes based on the distribution of those codes across other establishments with the same Standard Industrial Classification (SIC) and other characteristics. These procedures will be applied to current and, to the extent feasible, historic data on the LDB including establishments that are out of business. This provides a frame for surveys requiring stratification by NAICS and aids in the conversion from SIC to NAICS for ongoing surveys. In addition, the availability of a continuous history with NAICS codes will permit seasonal adjustment and other time-series analysis of the data.

2. WHAT IS NAICS?

NAICS was established in 1997 through a cooperative effort among the United States, Mexico, and Canada. The Bureau of Labor Statistics worked closely with the Bureau of the Census, the Bureau of Economic Analysis, and other U.S. statistical agencies to achieve the goal of developing NAICS. NAICS replaces the SIC (Standard Industrial Classification) system that has been in place since the 1930s and was last revised in 1987 (NAICS, 1998).

NAICS was developed based on the economic concept that establishments should be grouped together according to similar production processes. This coding system focuses on the identification of new and emerging industries and high technology industries, and provides increased detail in the services sector over what was available under the SIC system. It uses a six-digit classification system that generally provides three-country comparability at the five-digit level. Under NAICS, the highest level of aggregation is the sector, of which there are 21. This compares to the 10 divisions available under the SIC system. NAICS includes nine new service sector aggregations that were not found under the SIC system. For additional information concerning the NAICS coding system, see Ambler (1998) and Murphy (1998).

3. OTHER CHANGES WITH NAICS

The treatment of auxiliaries will change under NAICS. Auxiliaries are worksites within a company that primarily serve other establishments within the same company (examples are warehouses or corporate offices). Under NAICS, auxiliary units will carry the NAICS code for their primary activity, while under SIC, auxiliary units were classified according to the primary activity of the company they served. BLS is conducting a special survey in fiscal year (FY) 2000, in order to verify auxiliary status and assign NAICS codes to auxiliary units reflecting that status. Non-auxiliary units will be automatically assigned a code that matches their NAICS code. The success of the survey on auxiliaries will be especially important, because ES-202 data under NAICS will be tabulated and published using the NAICS treatment of auxiliaries.

A revision to NAICS 1997 is on the way -- NAICS 2002. The three countries are currently working on proposed changes to the construction and wholesale trade sectors, because agreements were not reached on these two sectors during NAICS 1997. In addition, changes are planned within the Information and Retail Trade sectors in order to better capture Internet-related activities.

The task for BLS and its State partners is to assign NAICS codes to the 8.2 million business establishments in the ES-202 program. At the same time, the ES-202 program will verify SIC codes (and assign SIC codes to new units) in order to create linkages between the two industry classification systems. BLS plans to implement NAICS over a four-year period. By the end of FY 2001, all establishments in the ES-202 program will be assigned NAICS 2002 codes. The first step in assigning NAICS codes will be to contact the employer directly, as described in the next section. Those units that do not receive NAICS codes through this process (i.e., nonrespondents) will be assigned one by an imputation process developed by BLS (described later in this paper). BLS will incorporate NAICS 2002 changes, using the NAICS treatment of auxiliaries, with the first publication of ES-202 data under NAICS. This will be published in 2002, for reference year 2001. This schedule will ease the burden on data users by providing a single change in coding structure from the 1987 SIC to NAICS 2002.

4. ASSIGNING NAICS CODES BY CONTACTING THE EMPLOYER

4.1 The Refiling Process

The ES-202 program updates classification codes using a process known as "refiling," in which the employer receives a form from their State Employment Security Agency (SESA). The employer will verify or update the information contained there, including the primary business activity of the establishment. The form will ask the employer to select an appropriate NAICS-based industry description for the establishment. The SESA will then assign a NAICS code based on this response. Some SIC and NAICS code combinations will be direct matches, that is, the SIC code is associated with only one NAICS code. Split combinations, or non-directs, occur when the SIC code maps to more than one NAICS code. During the refiling process, BLS targeted directs and non-directs, as well as records that had no NAICS code, an unclassified NAICS code, or an invalid one. During the last year of implementation, BLS will refile establishments affected by changes in NAICS 2002. Details are as follows:

FY 1998 Establishments in direct industries were automatically assigned a NAICS code by a computer program. This affected approximately one-half of establishments in the ES-202 program. The direct match program is run periodically to assign codes to any records that have direct match SICs but no NAICS code.

FY 1999 All units with employment greater than or equal to 50 (including directs recoded in FY 1998) were selected to receive a refiling form as well as units that had SIC 9999 (Unclassified), or SIC 9621, (Regulation and Administration of Transportation Programs). Records collected by the BLS EDI (Electronic Data Interchange) center were also refiled. Finally, a random sample of the UI accounts with less than 50 employees and worksites with SICs that could not be directly matched to one NAICS code were refiled.

FY 2000 Selected during this fiscal year were units that lacked a NAICS code, had an unclassified NAICS (NAICS 999999) or had an invalid NAICS code. A survey is being done in FY 2000 in order to verify the auxiliary status of auxiliary units and to assign a corresponding NAICS code.

FY 2001 Included in this year's refiling will be those SICs impacted by the NAICS 2002 revision.

4.2 Response rates

The success of the revision from SIC to NAICS requires that BLS and its State partners work diligently to ensure accuracy and completeness in the conversion to NAICS codes. To meet this objective, States pursue a goal of achieving usable response rates of at least 90 percent, in both units and employment, during each year's refiling cycle. Usable responses are those that receive a NAICS code through the refiling process. Establishments receive up to three non-response follow-up mailings. By December 1999, 72 percent of records in the ES-202 program had received NAICS codes from the refiling process; in terms of employment, 84 percent had been assigned NAICS codes. See industry details in the adjacent table. For the remainder of units that do not receive NAICS codes from the refiling process, BLS will assign NAICS codes through an imputation process as described later in this paper.

Division	Percent of Records Coded	Percent of Employment Coded
Agriculture, forestry, fishing	86.8	85.5
Mining	94.6	98.0
Construction	88.8	90.1
Manufacturing, durable	71.2	86.9
Manufacturing, nondurable	67.7	86.7
Transportation, public utilities	59.9	79.4
Wholesale trade	63.0	73.7
Retail trade	69.0	80.3
Finance, insurance, real estate	62.1	73.7
Services	72.7	85.5
Government	97.7	98.0
Total with NAICS	71.9	84.0

5. ESTIMATION WITH FIRST QUARTER 1999

Most establishment surveys that BLS conducts use historical time series data in order to evaluate current economic activity. The implementation of a new industry coding system has a significant impact on the continuity and value of these time series. Because the assignment of NAICS codes is phased in over a four year period, BLS programs that maintain time series need to be able to estimate the movement of economic activities between the SIC and NAICS codes before all of the establishments have been assigned NAICS codes.

For purposes of estimation, the records are divided into three types: direct matches, certainty records, and sample records. Each of these is handled separately during the estimation process. For the purpose of calculating weights used for estimation, UI accounts were stratified by state, 4-digit SIC, and employment size class.

5.1 Direct Matches

The direct matches are records that have only one NAICS code associated with the SIC code for that record. Included are single worksite accounts with an average monthly employment (AME) of 50 or less, and multiple worksite accounts with all worksites in direct SICs and a total AME of 50 or less. In the estimation process, these records receive a final weight of 1.000, and a non-response adjustment is not needed.

There is no accounting for out-of-business UI accounts although some proportion would fall into this category. Since these are small UI accounts, out-of-business units could be a substantial part of this category. This may lead to an overestimation of the number of units and employment.

There is also no accounting for movement of these direct units into non-direct NAICS codes or into other direct NAICS codes among these small employers. For example, all units in SIC 0112 would be coded to NAICS code 111160. They cannot be classified into NAICS codes like 111140 or any of the other non-direct NAICS codes like 111150. Therefore, the number of units and employment for direct NAICS codes would be overstated while for non-direct NAICS codes it would be understated.

5.2 Certainty Records and Sampled Records

We designated UI accounts with AME of at least 50 as "Certainty" and selected 100 percent. We randomly selected about one-half of the UI accounts consisting of single records in split SICs with an AME of less than 50, and multi-unit accounts that have at least one sub-unit in a split SIC, and an AME of less than 50. We called these "Sampled." Each of these groups of UI accounts, Certainty and Sampled, was stratified across 4-digit SIC and size class.

Within each stratum, we calculated sampling weights, *N/s*, where *N* is the number of UI accounts in a stratum and *s* is the number of accounts selected for refiling. The weight was generally 1.000 for certainty strata and about 2.000 for sampled strata. The non-response adjustment factor is *s/r*, where *s* is the number of UI accounts that were selected and *r* is the number of UI accounts that responded including out-of-business UI accounts. In the absence of any other information, the assumption is made that the distribution of non-respondents is the same as that of respondents.

For multi-establishment UI accounts, a partial response is considered a respondent. For these UI accounts, a weight adjustment is done to account for the non-responding sub-units. This adjustment, *p*, is the ratio of the sum of employment across all reporting units in the account divided by the sum of employment across all of those with NAICS codes. In addition, for multi-establishment accounts, all sub-units have to be out-of-business for the account to be classified as out-of-business. The final weight is equal to the sampling weight times non-response adjustment times the partial adjustment, $(N/r) * p$, and is assigned to each sub-unit of a UI account. Estimates were calculated by summing data of appropriate establishments to aggregated levels. Essentially, the formulas in the box were used. fwt_i are the final weights and the summation is across all reporting units in any group of interest. Ratio tables that show the distribution of units, employment, and wages from each SIC across the NAICS codes associated with the SICs were also produced.

$$\hat{N} = \sum_i fwt_i$$

$$E\hat{M}P = \sum_i (fwt_i)(EMP_i)$$

$$Wa\hat{g}es = \sum_i (fwt_i)(Wages_i)$$

6. IMPUTING NAICS CODES WHERE THEY ARE MISSING

We need NAICS codes assigned to every record in the database for sampling on a NAICS code basis and for aggregating records to publish summaries and other statistics. We will apply an imputation procedure, state by state, to assign NAICS, NAICS corresponding to auxiliary status, and NAICS 2002 codes where they are missing on the 2000 and 2001 files.[1] For an overview of imputation procedures, see Kalton and Kasprzyk (1982).

Our imputation of NAICS 1997 in the summer of 2000 and NAICS 2002 in the summer of 2001 will use a nearest neighbor procedure. First, we will assign NAICS codes automatically to records with direct match SICs. Then we will apply an imputation procedure to assign NAICS codes to any records that remain without them. This nearest neighbor procedure will choose a donor record with the closest average employment from among those records with the same SIC and a state-assigned NAICS code. Ties among donors will be broken with a random assignment process. The process is based on the assumption that among records with the same 4-digit SIC, employment is a significant explanatory variable when determining NAICS assignment. The algorithm will be applied first to records from UI accounts with multiple worksites reported, and then to any remaining records without NAICS codes. Before imputing codes, the files will be edited for invalid SIC/NAICS conditions. Records that do not pass this edit will not be used in the imputation process and will be forwarded to the states for correction.

[1] In order to approximate the distribution of NAICS codes across records with reported SIC and NAICS codes, the Bureau of the Census used a random assignment process. This process used digits from the Employer Identification Number (EIN) from each record missing a NAICS as random numbers. They established ranges for random assignment of NAICS based on the proportion of records assigned each NAICS in a given SIC/NAICS group (Census, unpublished internal memo). Statistics Canada did not have to deal with missing NAICS codes because they handle industry coding centrally.

Occasionally, none of the records for a given SIC will have a state-assigned NAICS. In these cases, the procedure will go to a national summary file, with records of observed SIC and NAICS combinations, and will choose a donor record with the closest average employment from among those with the same SIC, and assign that NAICS code.

The first type of record requiring imputation comes from UI accounts with multiple worksites, where some records in that account for a given SIC have NAICS codes reported but others do not. The imputation will be carried out using only records reported with that UI account. For each SIC assigned to records in any such UI account, we will determine whether any records have a NAICS code assigned. If none of the records has a NAICS code, then we will calculate the average employment across the records, search the national summary file and choose from among those with the same SIC the NAICS code with the closest average employment. If some of the records with the same SIC in the UI account have a NAICS code assigned, we will determine whether there is only one NAICS code or more than one. If there is only one NAICS code, then we will assign that code to every record with that SIC. Otherwise, we apply the nearest neighbor method to assign codes.

We will impute for any single-site UI account without a NAICS code using the same algorithm. First, we will attempt nearest neighbor imputation using responses from records with the same SIC. If no such record are available, we will search the national summary file and choose the NAICS code with the closest average employment from those with the same SIC. In the event that a particular SIC has no responses for NAICS in the national summary file, we will determine the possible NAICS codes for that SIC and assign those codes randomly across the UI account having that SIC.

The process of imputing NAICS codes that reflect auxiliary status where those codes are missing will treat records the same whether they come from a multiple-worksite UI account or a single-site account. Because these records will be so limited in number, the procedure will go directly to one of two national summary files. One file has all reported NAICS, and auxiliary code combinations with reported NAICS codes reflecting auxiliary status and average employment. The other file has all reported SIC and auxiliary code combinations with reported NAICS codes reflecting auxiliary status and average employment. First, any record without an auxiliary code or with a code indicating that the record is an operating facility will have the NAICS code assigned to the field for the NAICS code corresponding to auxiliary status. Records with an auxiliary code indicating that the record is a headquarters or regional office, will be assigned the NAICS code "551114" indicating it is a headquarters. For any other record with the NAICS and auxiliary codes reported, we will search the national file for the records with that NAICS and auxiliary code, and choose the auxiliary treated NAICS code with the closest average employment. If the record missing an auxiliary treated NAICS code is also missing the NAICS code, we will search records with the same SIC and auxiliary codes and assign the auxiliary treated NAICS with the closest average employment. If there is neither a NAICS nor an SIC code, we will search records with the same auxiliary code and assign the auxiliary treated NAICS with the closest average employment.

7. CREATING A TIME SERIES

The BLS maintains a longitudinal data base (LDB) that links UI reports from businesses through ownership changes, to the extent possible. Each quarterly record on the LDB has an LDB number that links the records for an establishment back through time. This makes historical data for the establishment easily available for analysis. Over time the industry classification of some establishments in the data base have changed. The ES-202 program is considering how to assign NAICS codes to historical records, in light of these classification changes. One approach would be to assign the code imputed for the most recent quarter to earlier quarters as well, regardless of changes in classification. An alternative approach would be to assign the NAICS codes independently each time the classification changes, applying the same algorithms used to assign codes to First Quarter 2001 records. That NAICS code would then be carried back in time as long as the classification remains unchanged.

Establishments that have gone out of business prior to First Quarter 2001 are also part of the LDB and historic time series, therefore they must be assigned a NAICS code based on SIC. These records will be assigned codes by applying the same algorithms within each quarter.

8. CONCLUSION

Implementing the NAICS at BLS will involve a multi-year process that is nearly complete. It is imperative that the implementation be done in both an accurate and timely manner. Other BLS programs will begin publication based on NAICS effective with data year 2001. (See the box.) Therefore, by data year 2001, all 8.2 million employers in the ES-202 program must have NAICS codes. BLS has produced estimates of the effect of NAICS on employment and wages reported under the ES-202 program. Normal nonresponse will require BLS to assign NAICS codes to those employers that remain without one after the refiling process ends. BLS has designed an imputation process to assign NAICS codes to those records that do not have them.

BLS IMPLEMENTATION SCHEDULE

Reference Year	Program
2001 -	Covered Employment and Wages--ES-202
	Job Openings and Labor Turnover Survey
2002 -	Occupational Employment Statistics
	Mass Layoff Statistics
2003 -	Current Employment Statistics
	Productivity measures for selected industries
	Foreign Labor Force Statistics
	Occupational Safety and Health Statistics
	Current Population Survey
2004 -	Employment Projections
	National Compensation Survey
	Producer Prices Indexes

9. REFERENCES

Ambler, Carole A. (1998), “NAICS and U.S. Statistics,” *Proceedings of the Section on Government Statistics and the Section on Social Statistics, American Statistical Association, pp. 21-30*.

Executive Office of the President, Office of Management and Budget (1998), *North American Industrial Classification System - United States, 1997*.

Kalton, G. and Kasprzyk, D. (1982), “Imputing for Missing Survey Responses,” *Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 22-31*.

Murphy, John B. (1998), “Introducing the North American Industry Classification System,” *Monthly Labor Review,* July 1988, Vol. 121, No. 7, pp.43-47.

U.S. Department of Labor, Bureau of Labor Statistics, Bulletin 2490 (1997), *BLS Handbook of Methods, pp. 42-47*.

THE TRANSITION FROM SIC TO NAICS – A DISCUSSION

Leonard M. Gaines, Empire State Development[*]
30 South Pearl Street, Albany, New York, 12245, U.S.A
lgaines@empire.state.ny.us

ABSTRACT

The four papers that were presented in this session are discussed. The paper by McMahon describes the problem that the other papers suggest alternate solutions to. Each of these solutions is based on assumptions that might or might not be valid. This discussion identifies these assumptions and suggests alternate methods of handling the major problems that might invalidate these assumptions. The discussion concludes with some general comments about these papers and suggestions for future research.

Key Words: Industrial Classification, Standard Industrial Classification, North American Industry Classification System

1. INTRODUCTION

Before I begin to discuss these papers in detail, I want to inform you of my background and the biases that result from it. I work in the Policy and Research Division of New York State's economic development agency. Like many of my colleagues here, part of my duties is to be an industry analyst. However, with the change in industry classification methods, I am not sure what industries I analyze. My duties lead me to be very interested in small area, really county-level, trends and the data needed to analyze those trends. Another bias that I have is that I am looking for the best data feasible for policy purposes. This involves a trade-off of cost, development time, and data accuracy improvement.

All of these papers fit together very well. The McMahon paper provides a good introduction to the basic problem that the other papers propose alternate solutions to. The problem that the three remaining papers attempt to resolve is how do you assign North American Industry Classification System (NAICS) codes to establishments based on limited information, especially when it is not feasible to contact each establishment directly to get additional information.

The three papers propose similar methods, detailed in each paper, for assigning codes to establishments in an individual Standard Industrial Classification (SIC) industry that transitions to only one NAICS industry and where there is sufficient information from the establishment to assign a NAICS code. The real problem is the establishment that does not meet either of the above conditions. Each of the three papers describes a different method for dealing with this problem.

This discussion is organized with a separate section containing my comments about each paper's method of responding to the unassignable NAICS code problem. It then concludes with a section containing my general observations about these papers and suggestions for future research to resolve some of the key questions left unanswered by these papers.

2. THE McMAHON PAPER

This paper demonstrates that changing industry classification systems does impact the quality of the data being produced by systems relying on industry as part of their sampling frame. This is especially true in those circumstances where historical industry trends are used as a criterion in designing the sampling structure. While these impacts are likely to exist for only one or two sampling design revision cycles, which can be up to five years if the cycle is based on the Economic Census, and are likely to be relatively minor impacts, data users need to be aware that they do exist.

[*] The opinions presented in this paper are solely those of the author and do not in any way reflect the opinions of Empire State Development or New York State.

McMahon really brings out the need to assign NAICS codes to establishments – or in his case, principle business activity (PBA) codes to companies – to observations where they are missing. Since over 15% of the sample has an unknown PBA, and this is the second largest category in the sample, it is possible that the industrial distribution of this group could impact the Internal Revenue Service's (IRS) sampling frame. In order to improve the sampling frame, IRS should have a mechanism for assigning valid PBA codes to companies lacking one.

In addition to developing a method for imputing PBA codes, areas for future research that I see related to this paper focus on corporations rather than partnerships. One research focus should look at the differences in missing PBA code patterns between corporations and partnerships. Another research area that should be considered is the impact of a corporation changing its PBA over time. This may be more of a problem with large corporations that tend to be involved in a number of different activities, such as a manufacturing company getting involved with financing the purchase of its own products and gradually earning most of its income from its finance business activities instead of manufacturing.

3. THE KLIMEK AND MERRELL PAPER

This paper describes the authors' attempt to create NAICS-based aggregate tabulations from the 1992 Economic Censuses. While this effort is one that is to be lauded, there are several problems with their work. The first is that it has limited usefulness because of it is currently limited to the trade sectors from the 1992 Censuses. Another limitation on the usefulness is the heavy reliance on national relationships as the foundation of the authors' imputation method.

The imputation method being proposed by the authors assumes consistent industry composition across the nation. In their paper they have identified two out of the three weaknesses of this assumption. The weakness that was missed is that there are likely to be regional differences in the SIC/NAICS mix in those SIC industries being split into more than one NAICS industry.

The authors of this paper propose several refinements to their imputation method. The first is to work with firm-level data where possible. This is a good improvement. The other refinements proposed by the authors, the use of multinomial logistic regression and bootstrapping, would make economists and statisticians happy, but might be too much effort for a small gain in reduced variability in the aggregate data produced to make any difference in policy decisions.

There are several other refinements to their methodology that Klimek and Merrell should consider. There is a real need to consider accounting for geographic differences in the industry splitting patterns. The other area that the authors need to think about is how they will handle auxiliary establishments as they expand their work to the other economic sectors, and expansion of this effort to the rest of the 1992 Economic Censuses is something that is needed.

4. THE KONSCHNIK, HANCZARYK, AND KORNBAU PAPER

These authors propose assigning NAICS codes to establishments missing them, but with valid SIC codes, based on the national distribution of establishment births for the three years prior to the 1997 Economic Census. Generally, this is a good method, but it does have a potential weakness.

That weakness is that it assumes that the mix of SIC splits into NAICS industries is consistent over the three-year period being used to allocate the establishments. This is probably a good assumption, except possibly in those industries going through rapid change. Unfortunately, these rapidly changing industries are the same SIC industries that are most likely to be split into more than one NAICS industry. It might be better to use only the most recent one or two years' distribution if there are a sufficient number of births on which to base the allocations.

Konschnik, Hanczaryk, and Kornbau also propose using additional information, such as location, company name, and establishment size to assist in assigning NAICS codes. I think this is a good idea since, as a general rule, the use of additional information produces better results than ignoring it.

In their paper, these authors also propose using other sources of industry assignments, such as commercially available industry directories. I would recommend that these sources be very carefully evaluated before they are actually used. The reasons being that many of these sources actually classify companies rather than establishments, this could be useful for IRS purposes, but not for the needs of the Census Bureau or the Bureau of Labor Statistics. Also, many of these sources work from very vague descriptions of the company's or establishment's activities that often differ from official criteria for assigning these codes.

5. THE MIKKELSON, MORISI, AND STAMAS PAPER

This paper suggests using a nearest neighbor approach to assigning NAICS codes to establishments where they are missing. The authors propose using a state-by-state SIC/NAICS split distribution for most establishments and a national distribution for auxiliaries and those cases where there is no state data available. The assignment is done by averaging the employment of all establishments with missing NAICS codes in a company in a single SIC in a state. Then all of the establishments that are missing NAICS codes are assigned the NAICS code from the establishment with an employment that is closest to their average employment.

Generally, this is a reasonable approach but I am concerned about averaging the employment across all establishments that are missing NAICS codes. The reason for this concern is that this method assumes that all establishments in a company that are assigned to a single SIC do almost exactly the same thing. This may not be the case, especially where there is a large employment variation at the establishments in that company. It might be better to group the establishments into size classes and use the average employment of that class to assign NAICS codes or to use the nearest neighbor of each individual establishment that has not been assigned to a specific NAICS industry.

For example, one company in New York State reported establishment-level employment that ranged from 1 to 443 at 25 establishments within a single SIC industry that is being split into more than one NAICS industry for an average employment of 19 employees. All but one of these establishments employed between 1 and 21 people. That remaining establishment employed 443. It is very possible that one establishment's activities, and that of some of the much smaller ones, would place it in a different NAICS industry than the others. Given the proposed imputation methodology any of these establishments that were missing NAICS codes would be assigned to the same NAICS industry.

As mentioned above, these authors propose using the national distribution within the industry, without regard to the company, to assign a missing NAICS code to auxiliary establishments. It would probably be better to include some company-based information to avoid the situation of a single company having two different administrative auxiliaries being assigned to the corporate management NAICS industry. An even better approach would be to send these back to the state and ask the state partner to assign the NAICS industry on the best information available to them, which would include anecdotal knowledge probably held by the local labor market analyst.

Finally, these authors propose linking the records for a particular establishment back through time in order to assign a NAICS industry in the past and develop a time series. This is an effort that is essential for trend analysis. However, care needs to be taken to avoid the assigning NAICS industries so far into the past that the industry did not exist. For example, given that the activities of retail establishments gradually changes over time, they need to be careful that one of today's video rental stores is not assigned to that industry in the late 1970s, before the videotape player existed in the residential consumer market.

6. CONCLUSIONS

Generally, the methods of assigning NAICS industries in the cases where they are missing proposed by the three papers in this session dealing with that problem are all very good. However, there are different potential weaknesses in the assumptions underlying each of the proposed methods.

While I have proposed possible solutions to overcome or lessen the impacts of these assumptions, it is impossible to determine the true impacts of these assumptions or proposed remedies without further research. Without this additional research it is also impossible to determine which of these methods is most likely to perform best in assigning the appropriate NAICS industry to those establishments that are missing it.

I suggest that in order to answer these questions, these authors collaborate, to the extent that they legally can, to conduct this evaluation. Specifically, they should compare their different methods, along with the proposed improvements, to a set of data containing artificially missing NAICS codes.