Updated unified category system for 1960-2000 Census occupations

Peter B. Meyer

Office of Productivity and Technology[1]

U.S. Bureau of Labor Statistics

November 1, 2009

Preliminary and incomplete

This document does not represent views or policies of the US Department of Labor

Abstract

An earlier paper proposed a consistent category system for occupations in the U.S. Census of Population data from 1960 to 2000, based mainly on the 1990 Census occupation definitions. This paper corrects and updates that work based on several sources including special data sets from the Bureau of the Census and the Bureau of Labor Statistics. In the previous assignment, all persons assigned to each occupation category in each year were kept together when assigned to the standardized category system. In this revised assignment scheme, some respondents are reassigned probabilistically, based on other information about the respondent such as their industry of work or employer. This uses more detail and is more accurate.

1. Concepts and definitions

The decennial U.S. Census of Population provides data on the earnings and occupations of individuals living in the United States. Occupations are recorded as a three-digit number matching one of several hundred job title categories defined for each Census by the Census Bureau. The lists change each decade. For a variety of reasons, researchers may want to use an occupational category system that is stable over time in the various data sets that use some version of the Census categories, including the decennial Census of Population data, the CPS, or the NLSY.[2] The Minnesota Population Center[3] created such a system. To any respondent since 1850 who had reported an occupation, they assigned one of the 287 occupations in from the 1950 Census occupation list and report this in their *occ1950* variable.

For a variety of reasons a researcher may want a more recent or more detailed category system than the 1950 Census one. In an earlier paper (Meyer and Osborne, 2005) all occupations in the 1960 Census or later were mapped to a more detailed list based roughly on the 1990 Census list. The unified list is a coarsened version of the 1990 list, in which the 504 occupations of the 1990 list have been combined into 289 categories. The 1980 occupation list was very similar to the 1990 list, which makes much of the time series simpler. The coarsening was helpful in cases where the 1960 or 1970 lists looked very different from the 1980 list and it was not feasible to include all the detail of the 1980 list and go back before 1970. For more detail please see the 2005 paper.

---

[1] With thanks to Anastasiya Osborne, Trent Alexander and others at IPUMS, David Autor, Michael Handel, and colleagues in the BLS Office of Employment and Unemployment Statistics for data, advice, and valuable comments. Judy Yang and Jon Bernt provided very valuable research assistance. Views and findings in this research do not represent official views, findings, or policy of the U.S. Bureau of Labor Statistics.

[2] CPS stands for the Current Population Survey, a monthly survey of the U.S. population, and the NLSY stands for the National Longitudinal Survey of Youth.

[3] The project is known as IPUMS, which stands for Integrated Public Use Micro Samples. The ongoing project is discussed at http://www.ipums.org. For more, see Ruggles and Sobek (2003), and King, Ruggles, and Sobek (2003).

The purpose of this paper is to show some small improvements to that "occ1990" set of assignments and to build on the set of tools available for others to build such harmonized classifications by imputing occupation from other information about the respondent such as income, age, sex, and years of education.

For many research questions, the best possible accuracy and precision in the occupation category is useful. A test of a particular hypothesis may require more detailed occupations for comparison, or larger subgroups in order to provide larger samples to generate reliable summary statistics for each group, such as the variance of earnings. Also, the researcher may wish to study a panel of occupations to see how computer technology has affected various occupations in the U.S., or the effect of changes in licensing requirements or unionization on some occupations. Over time it becomes more difficult to match new occupations to the 1950-based classification.

In Meyer and Osborne (2005), starting from the 1990 Census occupation list, we combined several detailed occupations into more general categories (making the occupation set more coarse) in order to provide a consistent time series for other Census years. We ended up with 389 occupation categories, some of which were special cases because they exist only in the 1960 data, or are "unknown" or "unemployed." Others were rough groupings, lumping some categories together to make stable long-term categories.

Combining categories requires some judgment about what the occupation classification should accomplish. Occupations are often distinguished from one another mainly by the kinds of *tasks* the workers perform. Sometimes they are defined based on the *function* the workers provide for others, or by the *hierarchical relation* between the worker and others (e.g. supervisors and apprentices). When occupations are organized by function, i.e. the type of service provided to other people, instead of by task, technical change tends to occur within occupational categories without altering occupation classification. For example, the work of nurses has adapted to technological change, but the occupation category "nurses" has remained consistently defined. Less often, technological change creates or wipes out categories. For example, the blacksmith occupational category existed in the Census classification until 1970, but not later. A category for computer scientists first appeared in the 1970 Census.

This occupational classification system was meant to support a study of high tech occupations over time, and it was preferred not to have categories appearing and disappearing but rather to have a long time series. So when there was a choice, we defined a new occupation category by the worker's function for others, not by task or hierarchy. E.g. when groups had to be combined, to the extent possible blacksmiths would be kept with other metal workers (rather than as a disappearing category) and apprentices and supervisors with the functional category, not separate.

Vast data is available in these categories, and small improvements in the assignments of occupations have the potential to be reflected in many studies. The analysis below was performed on the basis of 1% samples from the decennial Census of Population data for 1960-2000, downloaded from www.ipums.umn.edu. The CPS also used Census of Population occupational categories since 1968. The 1968-1970 March CPS used the 1960 Census occupation definitions, the 1971-1982 CPS data used the 1970 Census definitions, the 1983-1990 CPS apply the 1980 Census occupation categories, the 1991-2002 CPS data use the 1990 Census categories (with slight changes, documented on the IPUMS web site), and starting with the 2003 CPS the 2000 Census occupation definitions have been applied. The Census data offers large samples, but only every ten years, while the CPS has smaller samples of earnings and occupation data for every year.

The 2005 paper's category system has been used in research. With some changes (including an extension back to 1940 data) it was incorporated into the IPUMS *occ1990* variable which can be downloaded with population census data for the convenience of the user. In addition we have responded to a 44 requests for the program that creates this category system. In some cases this was by users of the NLSY (National Longitudinal Survey of Youth) data sets which in its earliest years used 1970-census occupations. A number of these users found mistakes. We received corrections, advice, and feedback from Philip Cohen, of Duke University's sociology department. We also studied the program written by Sarah Porter of the University of Iowa and discovered two more errors. Carol Scotese Lehr reported an error in Appendix C of the 2005 paper which will be updated soon. In other cases there was an misassignment because the original assignment was based what seemed to be the best-matching names of the occupation, when in fact the classifiers would have generally made a different choice. The major source of improvement explored here comes from special data sets described in the next section, which makes it possible to assign persons from a given source occupation to different harmonized occupations based on other information about the respondent.

2. Dual coded data sets

There are special data sets in which each respondent has been categorized by an official Census person (or other appropriate expert) into occupation categories from each of two different category systems. The occupation is then said to be "dual coded" or "double coded." With data like that we can study the micro information about the person that correlates to the way experts assigned them into categories. We are imputing 1980, 1990, and standardized occupations to the 1960 and 1970 and 2000 Censuses. There are two source datasets that make this possible, which also have micro data on the individuals so we can analyze which predictors in a source data set will predict how an individual would be categorized in a different category system. The coded or numeric variables available are approximately the same ones that were available to the Census classifiers – income, sex, age, years of education, and the class and industry. The Census classifier also saw, crucially, a description of the activities and tasks or job title, and a name or characterization of the employer, but we do not know of any way to get that information back.

The 1970-1980 Treiman file

A file sometimes called "the Treiman file" has 122,141 observations that were coded into the 1970 and 1980 occupation systems. It is kept at IPCSR. I received the data from David Autor of MIT. The variables which are approximately comparable between the Treiman file, the CPS, and the decennial Census include age, sex, race, income, and years of education; these can be used for imputation. According to Michael Handel of Boston University, the earnings in the Treiman file are from 1969.[4]

1990 to 2000

There is also microdata on individuals who were assigned both 1990-Census and 2000-Census occupations by the Census staff. CPS records for each month from 2000-2002 were coded into both classification systems. The combined data with our variables of interest is available from the author.[5]

One interesting advantage of the dual-coded sample is that because this sample includes data over time it is possible to incorporate changes over time in wages and probabilities so that one might incorporate time trends into the imputation of a 1990-vintage occupation to 2000s-decade data, or of a 2000-vintage occupation to the data from the 1990s.

3. Improved assignments from the 1970 data to the harmonized system

Personnel, training, and labor relations, specialists and managers

The 1970 category 56 is "Personnel and labor relations workers." In the dual-coded data set, most of these are mapped to either the 1980 category 8 "Personnel and labor relations managers" or to category 27, "personnel, training, and labor relations specialists." A sample of 414 from the 1970:56 category is available in the dual-coded data set, 89 of whom were categorized as managers, 322 as specialists. Another three respondents were categorized elsewhere and were dropped from the regression below.

---

[4] I intend to contact Donald Treiman, who made or used this data set extensively.

[5] Combining these records into one data set took some effort. The sources of that data were the monthly basic CPS data and related dual-coded add-ons found here: http://www.bls.census.gov/cps_ftp.html#cpsbasic_extract. That link goes to the dual-coded addendum, which is titled "CPS Basic Extraction 2000-2002" on the web page. To see the documentation of that dual-coded addendum, click on the link for the "2000 Based Public Use Extract Data Dictionary", and search for the names "NEIO1ICD" and "NTIO1OCD" to see the essential industry and occupation variables. These variables are in the 2000 coded system, at the most detailed level. Only the March monthly results will be usable.

Higher up on the web page are the main monthly CPS contents, which are coded in the 1990 Census categories. To see the documentation for these variables, see http://www.bls.census.gov/cps/basic/datadict/199801/puf98dd.htm.

One can match any observation of a person with an occupation in the basic-monthly-file from 2000 to 2002 to an observation in the dual-coded addendum, and vice versa, based on an exact match of all four of these variables: OCCURNUM, QSTNUM, year, and month.

It was then necessary to combine this data set with the March supplement to the CPS in years 2000, 2001, and 2002. I used many merge fields to match the March supplement to the other data set: the household (qstnum aka ph-seq), occurnum/alineo, age, sex, 1990 occupation, 1990 industry, education, survey-month, and survey-year. The resulting sample is big, with 32,494 from March 2000, 31,227 from March 2001, and 38,460 from March, 2002, totalling 102,181 respondents dual-coded.

We found good predictors of whether a worker was categorized as a "manager" or a "specialist" (in table 1):

- The self-employed were almost all categorized as specialists not managers.
- Almost all those in the federal or state governments were categorized as "specialists." We did not know why. Possibly government managers in these areas were classified elsewhere.
- Employees of employment agencies (1970 industry 737) were relatively more likely to be specialists. These are mostly workers assigned to work temporarily with a client employer.
- Females were more likely to be classified as specialists.
- Respondents with higher incomes, more education, and who worked longer hours were more likely to be classified as managers.

Table 1. Predictors of 1980-category occupation for personnel and labor relation staff in Treiman data set

Number of observations: 411 Pseudo R-squared = 0.1356
Dependent variable is 0 for managers and 1 for specialists

	Coefficient	Std error	p-value
Age	.189	.090	.036
Age-squared	-.002	.001	.053
Age less than 21	2.23	1.61	.165
State or Federal government employee	-2.82	.731	.000
Ln(earnings)	-.759	2.24	.735
Ln(earnings) squared	.048	.130	.716
Years of formal education	1.102	.816	.177
Years of formal education, squared	-.033	.028	.236
Is female	-.350	.353	.322
Constant	-10.861	11.92	.362

The final algorithm was: Among those with 1970 occupation 56 (Personnel and labor relations workers) assign 1980 occupation 8 (manager) if a certain value computed from the respondent's information was larger than a threshold, otherwise assign 1980 occupation 27 (specialist). The code in Stata is:

```
 gen testindex=-10.860497+.1889168*age-.00201975*age2-.75856806*lnearnings+.04751241*lnearn2+2.2312849*
> agelt21+1.1020654*educ+-.03331698*educ_squared-.35020781*female-2.8239163*ind_govt;
  gen assigned1 = testindex > -.59713253;
```

In the dual-coded sample, this code assigns 301 of the 411 (73%) to the 1980 category that the Census staff assigned.

Research workers

The 1970 category 195 is "research workers, not [otherwise] specified." In the dual-coded data set, most of these are mapped to either 1980 category 19, which is "Managers and administrators, n.e.c." or to 1980 category 235, "Technicians, n.e.c." A sample of 124 of the 1970:195 category is available in the dual-coded data set, 93 of whom were categorized in the managers category, 29 in the technicians category, and 2 categorized elsewhere. We found good predictors of whether one of these workers was categorized as a manager or a technician:

- The two self-employed workers were categorized as technicians.
- Higher income persons, male persons, more educated persons, and older persons were more likely to be categorized as technicians. This suggests that the manager and administrative category was not made up mainly of managers.[6]

The regression supporting this algorithm is analogous to the regressions in the previous examples and will be shared upon request. The resulting Stata code is:

```
gen testindex=-36.255442+.19518354*age-.00161001*age2+6.5641862*lnearnings
 -.39523704*lnearn2+2.1112275*agelt21+.4886623*educ+
 -.00579356*educ_squared+-.72569132*female;
```

[6] This is to be investigated: perhaps there is a high-income group of managers in occupation 8, or an outlier in occupation 27.

```
replace testindex=.0001 if emp_selfemp==1;
gen assigned1 = testindex > .66329422;
```

In the dual-coded sample, this code assigns 95 of the 124 (77%) to the 1980 category that the Census staff assigned.

Payroll and timekeeping work

The 1970 category 360 is "Payroll and timekeeping operators." In the dual-coded data set, all of these are mapped to either 1980 category 305, which is "Supervisors, financial records processing" or to 1980 category 338, "Payroll and timekeeping clerks." A sample of 289 of the 1970:360 category is available in the dual-coded data set, 260 of whom were categorized in the clerk category and 29 in the supervisor category. We found good predictors of how these workers were classified:

- The four self-employed workers were categorized as clerks.
- Higher income persons, male persons, and more educated persons were more likely to be categorized as supervisors.
- Persons under aged 21 were always categorized as clerks.

The regression supporting this algorithm is analogous to the regressions in the previous examples and will be shared upon request. The resulting algorithm is: Among those with 1970 occupation 360 (Payroll and timekeeping operators), assign 1980 occupation 305 (Payroll and timekeeping clerks) if a test index > -1.3862944, otherwise assign 1980 occupation 338 (Payroll and timekeeping clerks).

The code is:

```
  gen testindex=54.850838+.04594014*age-.00071951*age2-13.628952*lnearnings+.93633944*lnearn2+-2.0794569
> *educ+.09774936*educ_squared+.16647198*female;
  replace testindex=-10 if emp_selfemp==1;
  replace testindex=-10 if agelt21==1;
  gen assigned1 = testindex > -1.3862944;
```

That is, if the computed index is larger than the threshold, assign the supervisor job code 305, otherwise assign the clerk job code 338. In the dual-coded sample, this mechanism assigned 258 of the 289 (89%) to correctly, that is, to the 1980 category that the Census staff assigned.

Private household housekeepers and butlers

The 1970 category 982 is "Housekeepers, private household." In the dual-coded data set, these are mapped to either 1980 category 405, which is "Housekeepers and butlers" or to 1980 category 407, "Private household cleaners and servants." A sample of 196 of the 1970:982 category is available in the dual-coded data set, 81 of whom were categorized in the 1980:405 category and 115 in the 1980:407 category. These variables weakly predict how these workers in 1970:982 were classified in the 1980 system:

- Females and younger workers were more likely to be in the "private household cleaners and servants" category.
- Higher educated and higher paid workers were more likely to be in the "housekeepers and butlers" category.

The regression supporting this algorithm is analogous to the regressions in the previous examples and will be shared upon request. The resulting Stata code is:

```
gen testindex=-3.7667354+-.09946493*age+.00106499*age2
   -1.8870727*agelt21+1.7437463*lnearnings-.09516942*lnearn2
   -.36901586*educ+.02152176*educ_squared+-.5911348*female;
gen assigned1 = testindex > -.28185115;
```

If the test index is larger than the threshold, impute the first job code 1980:405 to the respondent; otherwise impute the second job code 1980:407. In the dual-coded sample, this mechanism assigned 61% of the respondents (120 of the 196) to the correct 1980 category that the Census staff assigned.

Cleaners, maids, and janitors

The 1970 category 902 is "Cleaners and charwomen." In the dual-coded data set, there are 772 of these. 430 were mapped to 1980 category 449, which is "Maids and housemen," and 326 were mapped to or to 1980 category 453,

"Janitors and cleaners." Another 16 were mapped to 1980:448, but for simplicity we will leave them out of the analysis. The following variables predict whether these workers in 1970:902 were classified in the 1980:449 category or the 1980:453 category:

- Females were far more likely to be put in the "maids and housemen" category than males were.
- Workers in the hospital industry were likely to be put in the "maids and housemen" category.
- Workers in the building services industry and workers younger than 21 were likely to be put in the "janitors and cleaners" category.
- Age, age-squared, log-earnings, log-earnings-squared and years of education were weak predictors.

The resulting Stata code is:

```
gen testindex= -11.639918+ -.03856515*age+ .00034859*age2+ -1.13523*agelt21+2.679698*lnearnings+-.1822503*
 lnearn2+.22103793*educ+-.01431282*educ_squared+4.2042363*female+2.0930025*ind_hospital
 -1.2078316*ind_buildingsvs;
 gen assigned1 = testindex > .66329422;
```

If the test index is larger than the threshold, we impute the job code 1980:449, otherwise we impute job category 453. In the dual-coded sample, this mechanism assigned 90% of the respondents (677 of the 756 used in the analysis) to the correct 1980 category that the Census staff assigned. The 16 others from the original 1970 category, who were assigned by Census specialists to the third category, 1980:448, will be incorrectly imputed one of these two. The overall accuracy is therefore 88%; this is the probability of assigning a person from the original large category correctly.

Summary

The five examples of imputation discussed above can be summarized in the table below.

Table 2. Summary of 1970-1980 dual-coded in-sample imputations

1970 category	dual-coded sample size	1980 categories	number in sample	predictors	in-sample accuracy
Personnel and labor relations workers	414	manager	89	higher income, more education, working longer hours, male	73%
		specialist	322	self-employed, fed or state govt, works for employment agency	
Payroll and timekeeping operators	289	supervisors, financial records processing	29	higher income, more educated, male	89%
		Payroll and timekeeping clerks	260	self-employed	
Research workers, not [otherwise] specified	124	managers, n.e.c.	93	female	77%
		technicians n.e.c.	29	self-employed, higher income, more educated, older	
Housekeepers, private household	196	Housekeepers and butlers	81	higher educated, greater salary	61%
		Private household cleaners and servants	115	female, younger	
Cleaners and	756	Maids and	430	female, in hospital industry	88%

charwomen	(of 772)	housemen			
		Janitors and cleaners	326	building services industry, age under 21	

The number of these imputations implied in the actual 1970 Census data can be computed. (to come)

Other potential examples

This technique can be used in more examples, and potentially in complicated ones. In one Census, some of the "athletes and kindred" category were physical education teachers. Possibly, teachers can be separated out because they worked in the public sector. There is a large "salespersons, not elsewhere classified (n.e.c.)" for which industry information should help split up some of the respondents into other categories. There is also a large "Foremen, n.e.c." category which existed in the 1960 Census, and we had to keep it in the proposed classification because there was no good category to match it to. This category can perhaps be split up by industry to align its members with the later categories which distinguished supervisors in extractive occupations from those in production occupations and several other categories.

3.1 Improved assignments from the 2000 data to the harmonized system

The 1990-2000 dual-coded data set has over 102,000 records but it has been difficult to find large occupations in which the predictors help the match to the 1990 categories. Here is one that worked.

Farm and ranch managers

The 2000 category 20 is "Farm, Ranch, and Other Agricultural Managers." In the dual-coded data set, there are 166 of these. Of these, 107 are matched to 1990 category 475, which is "Managers, farms, except horticultural," 21 are matched to 1990 category 479, which is "Farm workers," and the remaining 38 are matched to other occupations. (14 are matched to occupation 473 which may be worth investigating for predictors also.) These variables predict how these workers in 2000:20 were classified in the 1990 system:

- The self-employed were almost all put in the manager category, whereas almost all of the "farm workers" were employees of private firms.
- The four cases under age 21 were categorized as "farm workers." It seems reasonable to expect that this would hold in a larger sample also.
- Excluding those very young persons, older persons and persons with higher earnings were more likely to be categorized as managers.

The regression supporting this algorithm is analogous to the regressions in the previous examples and will be shared upon request. The resulting Stata code is:

```
gen testindex=-.78652605+.03267428*age+.39494719*lnearnings+-4.0155602*emp_private if !agelt21;
gen is475 = testindex > .34333333;
replace is475=0 if agelt21==1;
```

In the dual-coded sample, this mechanism assigned 69% of the respondents (114 of the 166) to the correct 1990 category.

Cost estimators

The 2000 category 60 is "Cost estimators." In the dual-coded data set, there are 96 of these. Of these, 40 are matched to 1990 category 22, which is "Managers and administrators, n.e.c.," 27 are matched to 1990 category 37, which is "Management related occupations, n.e.c.," and the remaining 29 are matched to other occupations. Persons in the construction industry were far more likely to be put in category 22, and higher earning persons were somewhat more likely to. It is not obvious why, but higher educated persons were more likely to be in category 37.

The regression supporting this algorithm is analogous to the regressions in the previous examples and will be shared upon request. The resulting Stata code is:

```
gen testindex=-2.2992037+-.60864667*educ+.84016698*lnearnings
             -1.1272302*emp_selfemp+4.7179597*ind_construction;
gen is22 = testindex > .30228087;
```

In the dual-coded sample, this mechanism assigned 60% of the respondents (58 of the 96) to the correct 1990 category, which though not high is more accurate than leaving them all in category 22. Remember that 29 of the 96 – over 30% could not possibly be matched by an algorithm that could only assign one of the two most likely ones.

Summary of the 1990-2000 imputations

2000 category	dual-coded sample size	1990 categories	number in subsample	predictors	in-sample accuracy
Farm, Ranch, and Other Agricultural Managers	128 (of 166)	Managers, farms, except horticultural	107	self-employed, older, higher income	69%
		Farm workers	21	employees of private firms; age < 21 ;	
Cost estimators	67 (of 96)	Managers and administrators, n.e.c.	29	in construction industry	60%
		Management related occupations, n.e.c.	27	more education	

Others will be possible but the job is not complete.

4.0 Extending inferences back to 1960

Filling out the actuaries in 1960

In the 1970 through 1990 Censuses, statisticians and actuaries were recorded as separate groups, but in the 1960 Census they were in one category, "statisticians and actuaries." In the earlier paper (Meyer and Osborne (2005)), when assigning 1990- based occupations to all the data from 1960 to 2000, we put the 1960 "statisticians and actuaries" into the statisticians category because it was much larger and therefore provides the closest match for most of them. We left the category for actuaries empty.

Table 3. Actuaries and statisticians in decennial Census samples

	1960	1970	1980	1990
Actuaries	199	45	129	182
Statisticians		237	352	338

Using other evidence about the respondents, we can infer which of them would have been likely to have been classified as actuaries in any later year, and move some of them into the empty 1960 actuaries category. Several predictors are pretty strong:

- Actuaries were much more likely than statisticians to work in the industries of insurance, accounting and auditing, or professional services
- Actuaries were less likely to work in government
- Actuaries were more likely to have high incomes
- Actuaries were more likely to have business income, in contrast to salary

- Residents of Connecticut, Nebraska, Minnesota, or Wisconsin were disproportionately likely to be categorized as actaries. These states have large insurance companies and related employment. Hartford, CT is an insurance center. Mutual of Omaha is headquartered in Nebraska. Blue Cross and Blue Shield has many employees in Minnesota.

- Actuaries were a growing fraction of the combined population over the years.

Using this kind of evidence, I ran exploratory tables and regressions to determine an accurate and feasible imputation of occupations to the 1960 subpopulation. This technique described below worked out well. I estimated a logistic regression (that is, "ran a logit") which predicts the probability that a particular respondent within this subpopulation is a statistician. Given a list of quantitative observations X_i for respondent i, and a set of coefficients β which will be estimated, this logistic function takes a complicated set of inputs and produces a value that is between zero and one which can be interpreted as a probability:

Pr(respondent *i* is a statistician) = Logistic($X_i\beta$) = $e^{Xi\beta}/(1+ e^{Xi\beta})$

This table shows the results of the logistic regression of these other variables in predicting which respondents were statisticians. The dependent variable is 1 if respondent was defined by the Census Bureau as a statistician, and 0 if the respondent was defined as an actuary. The independent variables in X_i are listed at the left. *Earned income* is defined here to be the sum of wage income and income from business or self-employment.

Table 4. Predictors of occupation for statisticians and actuaries in 1970-1990 Censuses

Number of obs = 1258 (1970-1990 Census, all statisticians and actuaries)
Pseudo R2 = 0.5333
Dependent variable is 1 for statisticians and 0 for actuaries

	Coefficient	Std error	p-value
year	0.074	33.139	0.000
Age	0.202	0.056	0.000
Age-squared	-0.002	0.001	0.001
Is in insurance industry	-3.818	0.284	0.000
Is in accounting/auditing industry	-4.775	1.158	0.000
Is miscellaneous services industry	-1.840	0.396	0.000
Is in nonprofit membership organization	-1.729	0.755	0.022
Is in professional services industry	-3.909	0.353	0.000
State government industry	-2.034	0.926	0.028
Ln(earned income)	-26.326	15.803	0.096
Ln(earned income) squared	2.881	1.566	0.066
Ln(earned income) cubed	-0.105	0.051	0.040
Fraction of earned income that is business income, not wages	-0.764	0.723	0.290
Years of education	-1.703	0.564	0.003
Years of education squared	0.046	0.017	0.006
Is classed as government employee	1.338	0.375	0.000
Is employed at time of Census	-0.659	0.403	0.102
Lives in Connecticut	-0.711	0.479	0.138
Lives in Minnesota	-1.191	0.724	0.100
Lives in Nebraska	-0.772	1.000	0.440
Lives in Wisconsin	-0.816	0.961	0.059
Constant	-51.805	66.446	0.436

Respondents with less than 13 years of formal education were never defined as actuaries. There were just a few of these. They were left out of this regression and assigned separately.

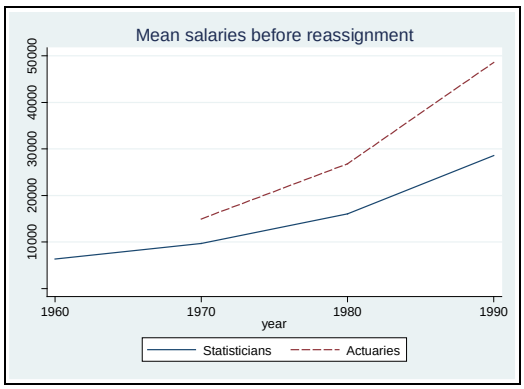
This evidence gives us the following algorithm to apply to the records in 1960 now categorized as statisticians, shown here in Stata code:

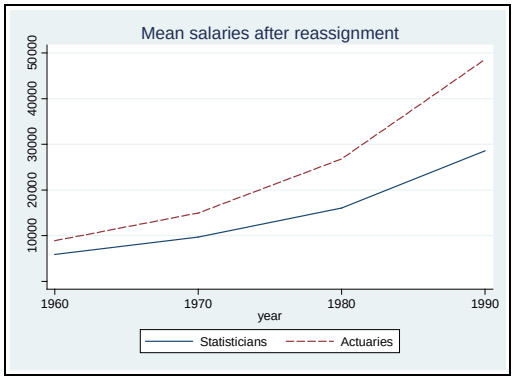
```
gen logitindex = 147.9366 * ln(year)
  + .2024399 * age
  -.0021747 * age * age
  -3.817868 * (ind1950==736)  /* 736 Insurance industry */
  -4.774511 * (ind1950==807)  /* 807 Accting and auditing */
  -1.840402 * (ind1950==808)  /* 808 Misc business services */
  -1.729038  * (ind1950==897)  /* 897 = nonprofit membership orgs */
  -3.909395    * (ind1950==899)  /* 899 = Miscellaneous professional and related
  -2.034102  * (ind1950==926)  /* 926 = state public administration */
  - 26.32612 * lninc           /* log(income) */
  + 2.880615 * lninc*lninc     /* income squared */
  -.1052547 * lninc*lninc*lninc     /* income cubed */
  -.7643481 * incbus / (incbus + incwage) /* fraction of business income */
 -1.702223 * educyrs
 + .0455556 * educyrs * educyrs
 + 1.338197 * govtemployee
 -.659389 * employed
  -.7113602 * (statefip==9)     /* lives in Connecticut */
  -1.190836 * (statefip==27)    /* in Minnesota, home of Blue Cross Blue Shield?
  -.772092 * (statefip==31)    /* Nebraska */
  -1.815364 * (statefip==55)    /* Wisconsin */
  -1026.72                     /* constant */
  ;
gen logitval=exp(logitindex)/(1.0+exp(logitindex))
replace logitval=.9999 if educlt13   /* this flag is a perfect predictor */
replace assigned = logitval>.45
```

The variable "assigned" then has a 1 for imputed statisticians, and 0 for imputed actuaries. The threshold value of .45 was found empirically to produce the right number of actuaries on the 1970-1990 data. That is, it misclassified equal numbers of actuaries and statisticians.

On the 1970-1990 data, this algorithm is 88% accurate. Let us assume that on the 1960, out-of-sample, data, it is 80% accurate. After the assignment, there are 30 actuaries. An estimated 24 new actuaries (80% of 30) were correctly assigned, and an estimated 6 of these newly assigned actuaries should have been statisticians, and an estimated 6 records left in the statistician camp are actually actuaries, but this problem is not made worse by the new assignment – they were already miscategorized.

Here are the mean incomes for the two groups, after the statisticians and actuaries were split in 1960.





Lawyers and judges

A similar situation occurs in the "Lawyers and judges" category. Lawyers and judges were combined into a single category in the 1960 data, but separate in all later years. The 2005 classification mapped them all into "lawyers" because this was the closest match. Only four or five percent of this category were judges in 1970-1990.

Table 5. Lawyers and judges and statisticians in decennial Census samples

	1960	1970	1980	1990
Lawyers	2053	2570	5082	7603
Judges		123	298	331

But in the 1970, 1980, and 1990 data, all judges worked in the public sector, and it may be possible to use information on the place of work (government versus other) to infer which of the respondents were most likely to be judges.

Within the lawyers and judges category, *all* of the private sector employees are categorized as lawyers. All judges report salary income, suggesting that an unemployed person was never defined as an unemployed judge, but rather an unemployed lawyer. Within the government sector, judges tended to be older and more highly paid, and were less likely to report any business income.

Here are results from a preliminary logistic regression analogous to the one for actuaries, restricted to those lawyers employed in the federal, state, or local governments because only these could possibly be judges, according to the 1970-1990 data:

Table 5. Predictors of occupation for lawyers and judges in 1970-1990 Census

Number of observations: 2659

(1970-1990 Census, all lawyers and judges employed in public sector)
Pseudo R-squared = 0.3392
Dependent variable is 1 for judges and 0 for lawyer

	Coefficient	Std error	p-value
Year	-.005	.010714	0.633
Age	0.155	0.033	0.000
Age-squared	-0.001	0.000	0.040
Federal government employee	-1.44	.137	0.000
State government	.499	.263	.058
Ln(salary)	-1.795	3.094	.562
Ln(salary) squared	.052	.333	.877
Ln(salary) cubed	.003	.012	.798
Ln(business income)	-.041	.036	.261
Fraction of earned income that is business income	-.714	1.053	.498
Education less than 16 years	2.235	.320	.000
Years of formal education	-.044	.046	.336
Is employed at time of survey	.224	.241	.352
Constant	13.017	23.428	.578

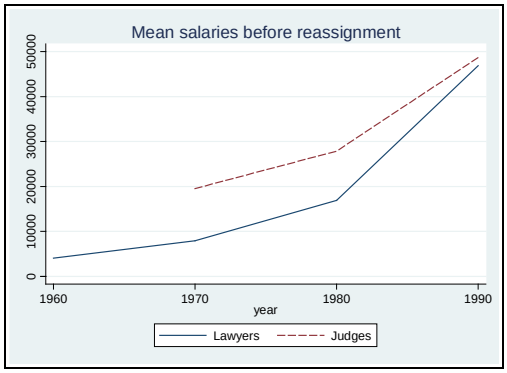
If one constructs a logistic index from the coefficients above, then applies the logistic function and reassigns 1970-1990 government-employed lawyers and judges with a resulting index of greater than .46 to be judges[7], the prediction is correct 84% of the time. On this basis, applying that same algorithm to the 1960 data we reassign 82 of the 2053 lawyers to be

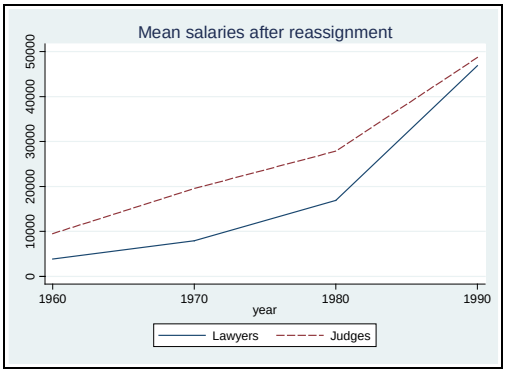
[7] Actually it is not necessary to use the logistic function here. The index itself must have a perfect threshold point at which the higher numbers are lawyers; the inverse-logit threshold of .46 would do it. A probit would work also, and it may be worth testing whether it gets the same results.

judges instead. Probably more of them were judges (see Table 4) but the assignments do not improve in accuracy and it seemed prudent to follow the Hippocratic principle by assigning as judges only the ones for which probabilities seemed highest.

Table 6. Percentage of lawyers-and-judges classified as judges

	1960 (imputed)	1970	1980	1990
Judges	3.99%	4.79%	5.86%	4.35%





5.0 Metrics of quality of the mapping

Sparseness of the resulting system

Are such efforts useful to those of us who are not specifically studying actuaries or judges? In a small way they are, because they improve the category system overall, allow more precise comparisons of other categories to a control group, and lengthen longitudinal panels and time series. Some categories are empty, however, when in principle they should not be because they are a best-match for some workers.

There are 295 empty cells in the 1960-2000 occupational categories if one uses the *occ1950* standard (with 287 categories, for each of five Censuses). Let us define a metric of the *sparseness* of the assignment to be the percentage of cells which are empty: 295/(287*5) = 20.56% of cells are empty. There are also 295 empty cells using our 2005 standard, which had 389 categories, so the sparseness metric is 15.17%. With the imputations for actuaries and judges, there are now 293 empty cells in this draft, or 15.06% by the sparseness metric. There are 155 empty cells from 1960, 82 from 1970, 6 from 1980, 5 from 1990, and 45 from 2000.

The same technique for imputing occupation can be applied more effectively on the 2000 Census categories because we have a set of dual-coded records available, in which the same records were assigned by the occupational coding specialists to both 1990 and 2000 Census occupations. Thus the imputation can be done with more confidence on the full 2000 data set. The data are hard to manage given our tools and we have no results for this yet, but the sparseness of this occupational category system can be driven down further by using such imputations however.

6.0 Conclusion

With an occupation category system lasting from 1960 to the present and large samples like those in the Census and CPS, researchers can test which attributes of an occupation predict other attributes of an occupation. For example, Meyer (2001) tested how an attribute of an occupation – the level of earnings dispersion within it -- evolved over time in particular types of occupations. The hypothesis was that high tech occupations and media-amplified occupations (called "superstars" occupations by Rosen (1981)) exhibited rising inequality within them.

One might extend this idea to treating each occupation as a separate labor market. This would give much more scope and precision to labor market theories and estimation. New Zealand's Department of Labor measures job vacancy rates by occupation, a practice which supports such an approach.

Another set of applications would treat attributes associated with occupations as predictors about individuals. For example, particular occupations have been identified as involving care work, very new technology, superstars' properties, and government licensing requirements. England, Budig, and Folbre (2002) tested whether caring and nurturing occupations (a gendered attribute) predicted pay levels apart from whether the jobholder was male or female. There is also a literature on the economics of income inequality, which could use narrow occupational categories as measures of skills.

A third set of applications to the methods proposed in this paper is to construct analogous long-lasting category systems for the industry variable in the Census and CPS. This would make it easier to identify long run trends in particular industries.

7.0 References

Advisory Panel on the Dictionary of Occupational Titles. 1993. Known as "the APDOT report." Downloaded from http://www.onetcenter.org/dl_files/PDF/AppendixC.pdf

Autor, David H., Frank Levy, and Richard J. Murnane. 2003. The Skill Content of Recent Technological Change: An Empirical Exploration. *Quarterly Journal of Economics* CXVIII: 4 (Nov, 2003).

England, Paula, Michelle Budig, and Nancy Folbre. 2002. Wages of Work: The Relative Pay of Care Work. *Social Problems* 49:4, pp. 455-473.

King, Miriam, Steven Ruggles, and Matthew Sobek. *Integrated Public Use Microdata Series, Current Population Survey: Preliminary Version 0.1*. Minneapolis: Minnesota Population Center, University of Minnesota, 2003.

Meyer, Peter B. 2001. *Technological uncertainty and earnings dispersion*. Northwestern University, Department of Economics dissertation.

Meyer, Peter B. Technological uncertainty and superstardom: two sources of changing inequality within occupations. http://econterms.net/pbmeyer/research/pdf/micro.pdf

Meyer, Peter B., and Anastasiya Osborne. 2005. Proposed Category System for 1960-2000 Census Occupations. U.S. Bureau of Labor Statistics working paper WP-383. http://www.bls.gov/ore/abstract/ec/ec050090.htm

National Crosswalk Service Center: http://www.xwalkcenter.org/

Rosen, Sherwin. 1981. The Economics of Superstars. *American Economic Review* 71:5 (Dec., 1981), 845-858.

Steven Ruggles, Matthew Sobek, Trent Alexander, Catherine A. Fitch, Ronald Goeken, Patricia Kelly Hall, Miriam King, and Chad Ronnander. *Integrated Public Use Microdata Series: Version 3.0* [Machine-readable database]. Minneapolis, MN: Minnesota Population Center [producer and distributor], 2004. Online at: http://www.ipums.org.

Scopp, Thomas M. The Relationship between the 1990 Census and Census 2000 Industry and Occupation Classification Systems. U.S. Census Bureau Technical Paper #65. Oct 2003. Online at: http://www.census.gov/hhes/www/ioindex/pdfio/techpaper2000.pdf

U.S. Department of Labor, Employment and Training Administration. 1991. *Dictionary of Occupational Titles*, fourth edition.

U.S. Department of Labor. 1993. *Labor Composition and U.S. Productivity Growth, 1948-90*. (pp 77-78 on substitute income for topcoded incomes)

U.S. Department of Labor. 1999. *Report on the American Workforce* chapter 3, "Economic change and structures of classification." http://www.bls.gov/opub/rtaw/chapter3.htm