Updated unified category system for 1960-2000 Census occupations

Peter B. Meyer

meyer.peter@bls.gov

Office of Productivity and Technology[1]

U.S. Bureau of Labor Statistics

January 2010

Abstract

Occupation categories in the U.S. Censuses of Population change every ten years. Meyer and Osborne (2005) proposed a standardized category system to report consistent detailed occupational data from the 1960 to 2000 Censuses, based on the occupations originally reported. This paper explores imputation of standardized occupation on the basis of further information about the respondent and special data sets from the Bureau of the Census and the Bureau of Labor Statistics which apply two Census occupation category systems to samples of respondents in the Current Population Survey. This technique uses more detail and in particular situations can be more accurate on average.

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 IPUMS project at the Minnesota Population Center created such a system.[3] 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. Users may download this variable along with the Census data.

For a variety of reasons a researcher may want a more recent or more detailed category system than the 1950 Census one.[4] Starting with the 1990 Census occupation list, Meyer and Osborne (2005) combined several detailed occupations into more general ones, making the occupation set more coarse, in order to provide a consistent time series for the Census occupations from 1960 to 2000. The unified list is a coarsened version of the 1990 list, in which the 504 occupations of the 1990 list are combined into 389 categories. Some of the categories were special cases that exist only in the 1960 data, or are "unknown" or "unemployed." Others were rough or awkward groupings but could last throughout the period.

The 1980 occupation list was very similar to the 1990 list, which simplifies the time series. The coarsening was helpful in cases where the 1960 or 1970 lists differed sharply from the 1980 list so it was not feasible to include all the detail of the 1980 list on earlier data.

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

[1] With thanks to Anastasiya Osborne, Trent Alexander and others at IPUMS, David Autor, Michael Handel, Jay Stewart, 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 or the U.S. Department of Labor.

[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 IPUMS project is discussed at http://www.ipums.org, and is cited as Ruggles and Sobek (2004), and King, Ruggles, and Sobek (2003). My understanding is that Matthew Sobek is the main author of the *occ1950* variable.

[4] There is no perfect category system for all occupation research. For many research questions, the finest possible detail on occupation is useful, whereas for others call for larger samples within every occupation category (e.g. for reliability of summary statistics) and therefore fewer occupation categories are better. In most systems there are various margins on which the assignment of one occupation rather than another is a judgment call.

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 to allow comparison of these occupation to long term consistent time series for other (control) occupations. 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 category which disappeared in the 1970s), and apprentices and supervisors were kept with their functional category (like electricians or plumbers), not separately by rank.

Vast data are available in the Census categories, and small improvements in the assignments of occupations have the potential to be reflected in many studies. The CPS has 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 use the 1980 Census occupation categories, the 1991-2002 CPS data use the 1990 Census categories (with slight changes during that period, documented on the IPUMS web site), and starting with the 2003 CPS the 2000 Census occupation definitions have been used. 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 at the convenience of the user. In addition we responded to dozens of requests for the program that creates this category system. Some of these requests were from users of the NLSY (National Longitudinal Survey of Youth) data sets which in its earliest years used 1970-census occupations. A number of users sent corrections, advice, and other feedback.

This paper explores how respondents might be better matched to one of these *occ1990* categories by imputing an occupation based on other information as well as the respondent's originally reported occupation. Section 2 applies a special technique in two ideal cases where one can stretch an imputation from 1970 back to 1960 because one 1960 category was split into exactly two 1970 categories with no ambiguity. Section 3 describes the dual-coded data sets that can be used on 1970 and 2000 data, which show empirical relationships between particular occupations and the respondent's industry, age, sex, education, years of education, and employer type (public, private, nonprofit, self-employed). Section 4 uses some resulting imputations on data from the 1970 Census, and section 5 applies the technique to data from the 2000 Census. The results are that this can indeed work sometimes; the technique is demonstrated to work on several occupations. Further work is needed to apply it fully and produce a useful improved *occ1990* variable.

2. 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.

The analysis below was performed on 1% samples from the decennial Census of Population data for 1960-1990, downloaded from www.ipums.umn.edu.

Table 1. 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 most likely to have been classified as actuaries in the later Censuses, and impute the occupation actuary to them in 1960, filling in a category that would otherwise be empty. There were just a few respondents with fewer than 13 years of formal education, and these were always labeled as statisticians not actuaries. That imputation rule is straightforward. Among the other respondents, there were several strong predictors:

- Actuaries were much more likely than statisticians to work in the industries of insurance, accounting and auditing, or professional services.
- 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. (Aetna is near Hartford, Connecticut. Mutual of Omaha is headquartered in Nebraska. Blue Cross / Blue Shield have many employees in Minnesota.)
- Actuaries were a growing fraction of the combined population over the years.
- Statisticians were more likely to work in government.

Based on these predictors, I ran exploratory 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 ("ran a logit") to predict the probability that a particular respondent within this subpopulation would be categorized as 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 estimates 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 2. Predictors of occupation for statisticians and actuaries in 1970-1990 Censuses

1258 observations (all statisticians and actuaries in 1970-1990 Census samples)
Dependent variable is 0 for actuaries and 1 for statisticians. Logistic regression. Pseudo $R^2 = 0.5333$

	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
Works in insurance industry	-3.818	0.284	0.000
Works in accounting/auditing industry	-4.775	1.158	0.000
Works miscellaneous services industry	-1.840	0.396	0.000
Works in nonprofit membership organization	-1.729	0.755	0.022
Works 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
Business income / (Salary + Business income)	-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
Works in government (according to *empclass*)	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

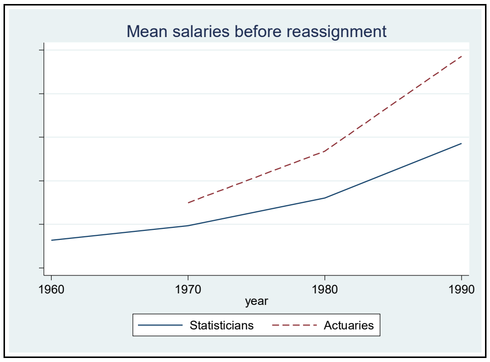
This evidence gives us the following algorithm to apply to the records in 1960 now categorized as statisticians, shown here in Stata code:

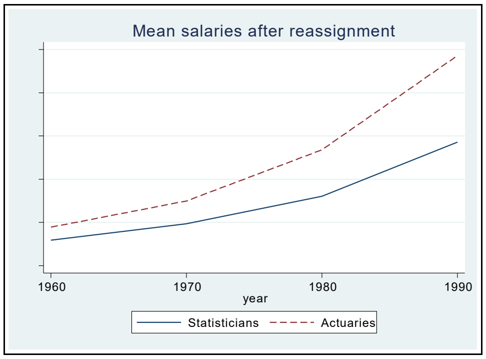
```
gen logitindex = 147.9366 * ln(year)
  + .2024399 * age
  -.0021747 * age * age
  -3.817868 * (ind1950==736)  /* 736 Insurance industry */
  -4.774511 * (ind1950==807)  /* 807 Accting and auditing */
  -1.840402 * (ind1950==808)  /* 808 Misc business services */
  -1.729038 * (ind1950==897)  /* 897 = 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 average incomes for the two groups in the Census samples before and after applying the imputation to the 1960 data. The point here is only that the imputed 1960 data points look reasonable.





Lawyers and judges

A similar situation occurs in the single 1960 "Lawyers and judges" category. Lawyers and judges were separate categories in the later decades. 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 3. Counts of lawyers and judges and statisticians in decennial Census samples

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

Several attributes in this pool predict whether the individual is categorized as a judge. All judges in the 1970-1990 data were employed in government, so for respondents who report a private employer one can immediately impute "lawyer." 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 logistic regression analogous to the one for statisticians and actuaries. The samples are restricted to those lawyers and judges employed in the federal, state, or local governments because only these could possibly be judges, according to the 1970-1990 data:

Table 4. 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)
Dependent variable is 1 for judges and 0 for lawyer. Logistic regression. Pseudo R-squared = 0.3392

	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

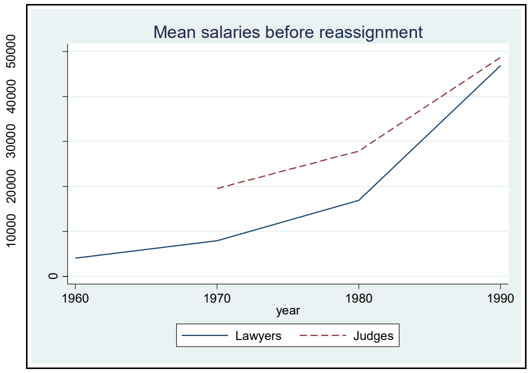
If one constructs a logistic index from the coefficients above, then applies the logistic function and reassigns 1970-1990 government-employed lawyers and judges with a resulting index of greater than .46 to be judges[5], the prediction is correct 84% of the time. On this basis, applying that same algorithm to the 1960 data we reassign 82 of the 2053 lawyers to be judges instead. Probably slightly more than this were judges (based on Table 5) but in the sample the imputed assignments on the 1970-1990 data do not improve in accuracy beyond this point. It seemed prudent to follow the Hippocratic principle by assigning as judges only the individual for which probability seemed to be quite high, otherwise the reassignment is very likely to create more errors than it fixes, even though the resulting number of judges matches one's best guess as to how many there should be. 82 imputed judges is also a sample large enough for aggregate statistics to be compared plausibly to other occupations.

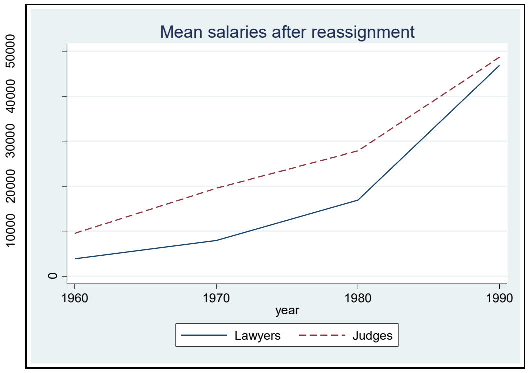
[5] Actually it is not necessary to use the logistic function itself in imputation. Since the logistic function is monotonic, the index itself has a threshold point at which the higher numbers are lawyers; the inverse-logit threshold of .46 would do it.

Table 5. Percentage of lawyers-and-judges classified as judges after imputation

	1960 (imputed)	1970	1980	1990
Judges	3.99%	4.57%	5.54%	4.17%

Graphs of the resulting average wage and salary income for these imputed groups appears reasonable so if one were to do a regression of income on occupation it would produce plausible results with the new data.





3. Dual coded data sets

The rest of this paper examines imputation into the Censuses based on special samples, not other Censuses. There are special data sets in which each respondent has been categorized by official Census specialist coders into occupation categories from each of two different category systems. The occupation is then said to have been "dual coded."[6] 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" (type of employer) and industry. Crucially, the Census classifier also saw a description of the activities and tasks or job title, and a name or characterization of the employer, but that information is not available in the data sets.

The 1970-1980 Treiman file

A file sometimes called "the Treiman file" (for sociologist Donald Treiman) has data on 122,141 respondents that were coded into both the 1970 and 1980 occupation systems. It is kept at IPCSR. I received the data from David Autor of MIT. The variables that 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. The data come from about 1980 when the transition between systems was being made.

1990 to 2000 dual-coded data

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, and in the March data we have information about the respondent to build a substantial regression. The resulting sample has 32,494 respondents from March 2000, 31,227 from March 2001, and 38,460 from March, 2002, totalling 102,181 dual-coded records. The combined data with the variables used here is available from the author.[7]

[6] Occasionally the term "double-coded" is used for this, although double-coded term more often refers to the practice of having two coders apply occupations to the same set of same respondents then comparing the results.

[7] Combining these records into one data set took some effort. The sources of the data were the monthly basic CPS data

This dual-coded sample includes data over time (March 2000, March 2001, and March 2002) so that one might incorporate time trends of wages or probabilities into the imputation of a 1990-vintage occupation to 2000s-decade data, or of a 2000-vintage occupation to the data from the 1990s. This paper does not go that far.

4. Improved imputations 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 in the 1980 manager category, and 322 in the 1980 specialist category. Another three respondents were categorized elsewhere and were dropped from the regression below.

These variables were good predictors of whether a worker was categorized as a manager or a specialist:

- The self-employed were almost all categorized as specialists not managers.
- Almost all those in the federal or state governments were categorized as "specialists." It is not clear 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.

Using these variables a logistic regression analogous to those in tables 2 and 4 then predicts the probability that a respondent is classified in the manager category or the specialist category.

Table 6. Predictors of 1980-category occupation for personnel and labor relation staff in Treiman data set
Number of observations: 411 Dependent variable is 0 for managers and 1 for specialists
Logistic regression. Pseudo R-squared = 0.1356. Coefficient on constant not shown.

	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

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 identifier (called qstnum on one data set and ph-seq in the other), the individual identifier (called occurnum or alineo), age, sex, 1990 occupation, 1990 industry, education, survey-month, and survey-year.

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 assigned_occ = 27;
 replace assigned_occ = 8 if testindex > -.59713253;
```

In the dual-coded sample, this code assigns 301 of the 411 (73%) to the 1980 category that the Census staff assigned. This imputation fills in a previously empty cell in the standardized assignment, because there had not been anyone from the in the standardized occupation category 8 before this imputation.

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. Perhaps there is a high-income group of managers in occupation 8, or an outlier in occupation 27; it may be possible to figure out why with further study.

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;
  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. There were 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 the test index is large enough, otherwise assign 1980 occupation 338 (Payroll and timekeeping clerks).

The resulting Stata 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 are summarized in table 7. In the 1970 Census samples these occupations total about .6% of the population.

Table 7. 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 government, 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 charwomen	756 (of 772)	Maids and housemen	430	female, in hospital industry	88%
		Janitors and cleaners	326	building services industry, age under 21	

5. Imputed occupations from the Census 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.
- Among those older than 21, older and higher-earning persons 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 Census 2000 category 60 is "Cost estimators." In the dual-coded data set, there are 96 of them. 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. Higher educated persons were more likely to be in the “management related” 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=-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. Another 29 of the 96 could not possibly be matched by an algorithm that could only assign one of the two most likely ones.

Appraisers and assessors

The 2000 category 81 is “Appraisers and Assessors of Real Estate.” In the dual-coded data set there are 110 of these. Of these, 78 (71%) are assigned to 1990 occupation 254, "real estate sales occupations,” another 20 to 1990 occupation 5, “Administrators and officials, public administration” and the others are in other manager or administration categories notably including generic category 22, "Managers and administrators, n.e.c." Employment information strongly predicted which 1990 occupation these respondents were put into. All the self-employed and almost all of those in industry 707 (“Real estate”) were assigned the “sales” occupation, whereas almost all of those in the “Public finance activities” were put into the “administrators and officials” occupation. No regression was necessary, which made it more feasible to assign three different 1990 occupations, not just two.

These statements assign a 1990 occupation with 90% accuracy to the dual-coded sample. Accuracy would have been 71% if they were all assigned to the largest category.

```
  replace predicted=254 if (empclass==6 | empclass==7)   /* 100% accurate for 45 cases out of 110 */
  replace predicted=5 if (ind2k==938 & predicted==.)  /* 95% accurate; match for 19/20 */
  replace predicted=254 if (ind2k==707 & predicted==.)  /* (93% accurate; match for 26/28 */
  replace predicted=22 if (ind2k>700 & ind2k<800 & predicted==.)
                             /* matches 2/4 right and for the others it makes a close match */
  replace predicted=254 if predicted==.  /* matches another 7/13 */
```

Personal financial advisors

The 2000 category 85 is “Personal Financial Advisors.” In the dual-coded data sets there are 309 of these, 267 of which map into one of three 1990 occupations: “Financial services sales occupations,” “Insurance sales occupations,” or “Other financial specialists.” Industry predicts this assignment accurately. If one assigns all those in the industry “Insurance carriers and related activities” to the insurance occupation, all those in the industry “Banking and related activities” to the “Other” category, and the rest to the “Financial services sales occupation” this achieves 211/309 = 68% accuracy. If instead they had all been assigned to the one best 1990 match, the accuracy would be 58%. No regression was necessary, although the use of one could possibly help since income and self-employment status varied among the target categories.

Summary of the 1990-2000 imputations

The next table summarizes the imputations discussed in this section. In the 2000 Census samples these occupations total about .3% of the population.

Table 8. 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	
Appraisers and Assessors of Real Estate	101 (of 110)	real estate sales occupations	78	Self-employed or in "real estate" industry	90%
		Administrators and officials, public administration	20	In "public finance" industry	
		Managers and administrators, n.e.c.	3	In other professional or administrative industry	
Personal Financial Advisors	267 (of 309)	Financial services sales occupations	179	All except two industries	68%
		Other financial specialists	64	Banking industry	
		Insurance sales occupations	24	Insurance industry	

5. Quality of the mapping

Sparseness metric of the resulting system

Are such efforts useful to those of us who are not specifically studying actuaries or judges or other narrow categories? In a small way they are, because they improve the unified category system overall. The imputations enable slightly more accurate comparisons of an occupation category of interest to a control group of persons in other occupations.

They also lengthen longitudinal panels and time series of certain occupations. In the classification of the 2005 paper, some categories appear empty, such as the 1960 actuaries above, 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 the *sparseness* of the assignment be measured by be the percentage of cells which are empty: 295/(287*5) = 20.56% of cells are empty. Using the 2005 *occ1990* standard (which had 389 categories), there are 295 empty cells, so the sparseness metric is 295/(389*5)= 15.17%. With the imputations in this paper for 1960 actuaries, 1960 judges, and 1970 personnel managers, it would be possible to fill in some good matches for three of these, leaving a sparseness metric of 292/(389*5)= 15.01% -- a tiny but measurable improvement. After these imputations, there are 155 empty cells from the 1960 data, 81 from 1970, 6 from 1980, 5 from 1990, and 45 from 2000.

Bias and overfitting concerns

There are some risks in imputing occupation then treating it as real data. Researchers may use a person's occupation as a predictor of the person's income, sex, or industry of employment. Here, we have defined the occupation category partly based on those variables. In principle this can lead to fitting the evidence circularly and thus exaggerating or overextending the importance of a predictor. For example, the income levels or sex ratios in an occupation imputed for 1972 may be inferred partly from the values for these variables in data from 1979 when the dual-coded data set was made. Since each such predictor is only one of several variables used in these imputations, and since the imputations are limited in number by the number expected in each occupation based on other evidence, these dangers do not seem serious to this author. Judgment is required, and like occupational classification in general, imputation of occupation is craftwork as well as science. One defends against risks by using imputation in moderation and openly discussing imputation methods.

6. 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 uses is to 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.

Another 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. The industry variable was also dual-coded at the same time and in the same way as occupation, so it would be possible to conduct this sort of study to impute a standardized industry over the same decades. This could make it easier to identify long run trends in particular industries.

This paper demonstrates the feasibility of imputing occupations in a number of special cases. Much work remains to try these techniques on the full set of 1970 and 2000 Census occupations, and then to complete the computer programs that apply the imputations to the full sets of 1970 and 2000 Census samples. Only then will it be feasible to update the values of the *occ1990* variable in the IPUMS databases of the Census samples.

It will take further effort to apply these methods to impute standardized occupations to the annual Current Population Survey data. It will be necessary to take inflation into account when the data come from a different year than the dual-coded data set. For example, the imputation will call for using 1971 income as an independent variable for predicting the occupation in the 1980 category system, when the dual-coded data set on which it is possible to estimate the effect of income comes from about 1980. Similarly the 2009 data on income must be adjusted to use the 2000-2002 dual-coded sample algorithm. Only then will it be feasible to impute the occupations in historical CPS and NLSY data in general.

7. References

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., 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. http://www.census.gov/hhes/www/ioindex/pdfio/techpaper2000.pdf