Updated Unified Category System for 1960-2000 Census Occupations

FCSM Federal Committee on Statistical Methodology conference
Washington DC Convention Center,   November 2, 2009

Peter B. Meyer

Office of Productivity and Technology,
U.S. Bureau of Labor Statistics (BLS)

. . . with much advice from others.

This research does not represent the Bureau itself, only the named author.

BLS

U.S. BUREAU OF LABOR STATISTICS
1 2 5 Y E A R S

www.bls.gov

Census Occupational Classifications

- U.S. Bureau of Census staff makes a list of 3-digit occupation codes for each population census

- Vast survey data is available in these categories
  - E.g. SIPP, NLS, ATUS, ACS, decennial Census
  - And/or mapped from other classifications
  - Current Population Survey (CPS) uses them:
    - ▶ 1960 system from 1968-1970
    - ▶ 1970 system from 1971-1982
    - ▶ 1980 system from 1983-1991
    - ▶ 1990 system from 1992-2002
    - ▶ 2000 system from 2003-2009

➔ Time series of occs have breaks over the period

- Research studies may call for a standard
  - Even if it's not nearly as good as the original!

2005 baseline to improve on

Earlier working paper (Meyer and Osborne, 2005) defines a unified classification of 389 3-digit occupations codes for Census & CPS from 1960 to present

It was adapted from the 500+ categories in 1990 Census:
- 379 categories have same name or almost same as 1990
- 125 were eliminated to help harmonize with other years
    - Example to follow
- 19 categories have expanded (changed name or n.e.c. category)
- 3 categories added for 1960 data which doesn't fit in

IPUMS.org (U of Minnesota History department had defined *occ1950* for 1850-present and used this category system for *occ1990*

Tradeoffs in Classification Systems

Duration of categories versus precision of categories
blacksmith, database admin (shorter, more precise series)
electrical engineer (longer evolving series)

Number of occupations vs. sample size of each
Narrow distinctions may be of interest
Dental technicians
High tech occupations vs. other technical occupations
"Superstars" jobs like athletes and musicians
Licensed jobs

Conformity with other data vs. "sparseness"
many missing year-occ cells

Meaning of occupation
tasks, function, skills, background, social class

There is no perfect classification but there are tools and criteria for making suitable ones.

Example: hard standardization case

1970 occ code	1970 occupation title	1980 occ code	1980 component categories and codes	Estimated number in civilian labor force	% of 1970 category
284	**Sales workers, except clerks, in retail trade**	**263**	**Sales workers, motor vehicles and boats**	185,160	**37.06%**
		266	Sales workers, furniture and home furnishings	98,941	19.80%
		267	Sales workers; radio, television, hi fi, and appliances	76,674	15.35%
		268	Sales workers, hardware and building supplies	81,668	16.35%
		269	Sales workers, parts	39,120	7.83%
		274	Sales workers, other commodities	16,008	3.20%
		277	Street and door to door sales workers	2,082	0.42%

We combined these

Input from users and new data

- Corrections from users
- Imputations from dual-coded data sets
  - 1970-1980   "The Treiman file"
  - 1990-2000   from Census and BLS, created from 2000-2003 CPS
- Imputations extending from 1970 back to 1960
- Visit to Census classifiers

125 YEARS
BLS

Industry and occupation coding at Census Bureau's National Processing Center in Jeffersonville, IN

- Industry codes and occupations codes are assigned by the same group of people, at the same time for each respondent.

- Industry is almost always decided first.

- The people who do it are "coders"

- Procedures are carefully documented

Coders

- **Coders** follow carefully documented procedures from the Census headquarters

- Coders with two years of experience are expected to assign 94 codes an hour, with 95% accuracy (which is checked)

- If there is not enough information to assign industry and occupation codes by procedure, the case is forwarded electronically ("referred") to a **"Referralist"** (aka statistical assistant)

Referralists

- There were about 12 coders and 14 referralists in October 2006
- Referralists have been coders before and usually have 9+ years of experience
- I interviewed three referralists, and a supervisor
- The ones I met handled referrals from several surveys:
  - ▶ CPS, ATUS, SIPP, NLS, ACS
  - ▶ Not decennial Censuses
  - ▶ Other surveys on contract
  - ▶ All these use decennial Census occupation codes

125 YEARS
BLS

Information Available to Coders and Referralists

- "kind of work"
- "principal duties"
- employer name
- city and state ("PSU") of respondent's home (not workplace)

- industry, already coded
- industry type (manufacturing, service, other)
- years of education, age, sex
- not income, since Jan 1994

- Industry is normally coded **before** occupation.

- Tens of thousands of job titles are mapped to a code in a reference book they have, if industry also matches.

- Some cases may be "autocoded" by software and coder checks

Information available to referralist

Given employer name:

- Can match to Employer Name List (ENL), same as SSEL or Business Registry.
- Can look on the web for employer
- Can study SOC manual ("little red book")
- May look up employer in Dun and Broadstreet data

Given tasks/activities:

- May refer to giant 1991 *Dict of Occ Titles* 1991

- They try to make a coherent choice for industry and occupation together.

Problems faced by referralists

- Having to hurry
- Too little information from respondent
  - "Computer work" for "kind of work"
  - Exaggeration (example: dot com businesses)
  - Ambiguity, e.g.:
  - "water company" for industry or employer
  - "surveyor" occupation
  - "boot" vs "boat" in handwriting
- Referralists confer with each other routinely, but sometimes make different choices from one another
  - Randomness is intrinsic even with more data
- Technological change goes along with occupational ambiguity

Person 1 (continued)

27 **Industry or Employer** — *Describe clearly this person's chief job activity or business last week. If this person had more than one job, describe the one at which this person worked the most hours. If this person had no job or business last week, give the information for his/her last job or business since 1995.*

a. For whom did this person work? *If now on active duty in the Armed Forces, mark ☒ this box →* ☐ *and print the branch of the Armed Forces.*

Name of company, business, or other employer

b. What kind of business or industry was this? *Describe the activity at location where employed. (For example: hospital, newspaper publishing, mail order house, auto repair shop, bank)*

c. Is this mainly — *Mark ☒ ONE box.*

☐ Manufacturing?
☐ Wholesale trade?
☐ Retail trade?
☐ Other (agriculture, construction, service, government, etc.)?

28 **Occupation**

a. What kind of work was this person doing? *(For example: registered nurse, personnel manager, supervisor of order department, auto mechanic)*

b. What were this person's most important activities or duties? *(For example: patient care, directing hiring policies, supervising order clerks, automobiles, reconciling financial records)*

29 Was this person — *Mark ☒ ONE box.*

☐ Employee of a PRIVATE-FOR-PROFIT company or business or of an individual, for wages, salary, or commissions
☐ Employee of a PRIVATE NOT-FOR-PROFIT, tax-exempt, or charitable organization
☐ Local GOVERNMENT employee (city, county, etc.)
☐ State GOVERNMENT employee
☐ Federal GOVERNMENT employee
☐ SELF-EMPLOYED in own NOT INCORPORATED business, professional practice, or farm
☐ SELF-EMPLOYED in own INCORPORATED business, professional practice, or farm
☐ Working WITHOUT PAY in family business or farm

13

Imputing judges

- In 1960 Census, "lawyers and judges" were one category. In unified classification we call them all lawyers.
- We wish to impute which ones are judges
- Plan: pool the 1970, 1980, 1990 Census lawyers and judges employed in the public sector
  - ▶ Lawyers and judges were were separate categories then
- Run logit regression for probability an individual is a judge
- Predictors
  - ▶ Older
  - ▶ Employed in state government
  - ▶ High salary income; low or no business income

We now impute judgeships

```
gen logitindex = -.0046652 * year + .1549193 * age
 -.0006942 * age * age  -1.4405086* indfed
 +.4986729 * indstate   -1.795481 * lnwage
 +.0517015 * lnwage * lnwage  +.0030016  * lnwage * lnwage * lnwage
 -.040749 * lnbus  -.7140285 * busfrac  +2.234934 * (educyrs<16)
 -.0442429 * educyrs  +.2239105 * employed      +13.0172   /* constant */ ;
gen logitval=exp(logitindex)/(1.0+exp(logitindex))

replace logitval=.0001 if !govtemployee   /* this is a perfect predictor */
replace logitval=.0001 if !indfed & !indstate & !indlocal    /* this too */

gen imputejudge = logitval>.46 /* Now 'imputejudge' has a 1 for imputed judges */
```

Threshhold probability is chosen to match the number of judges expected to be there, based on annual trend.
Predictions using this rule are **83%** accurate in the 1970 Pop Census.

After imputation:				
Census sample:	**1960**	**1970**	**1980**	**1990**
Lawyers	**1971**	2570	5082	7603
Judges	**82**	123	298	331

Statisticians and Actuaries

❑ Separate categories in and after 1970, but in 1960 they were all in "statisticians and actuaries" category

❑ When standardizing we put all these in "statisticians"

Counts of Actuaries and Statisticians in Census Samples				
	1960	**1970**	**1980**	**1990**
Actuaries	.	45	129	182
Statisticians	199	237	352	338

Will infer which people in this population were actuaries

Statisticians and Actuaries

- Pooled all 1970-1990 statisticians and actuaries
- Ran many logistic regressions predicting the actuaries
- Good predictors of whether respondent is an actuary:
  - Recorded in a later year
  - Employed in insurance, accounting/auditing, or professional services
  - Employed in private sector
  - High salary income
  - High business income, or mostly business income
  - Is employed
  - Lives in Connecticut, Minnesota, Nebraska, or Wisconsin (location of major insurance companies)

1960 Statisticians and Actuaries

- A 1970-1990 logistic regression predicts occupation **88%** of the time in 1970 sample.

- Applying it to 1960 data:

Revised counts of actuaries and statisticians after imputation				
	1960	**1970**	**1980**	**1990**
Actuaries	29	45	129	182
Statisticians	170	237	352	338

Mean salaries before reassignment

Mean salaries after reassignment
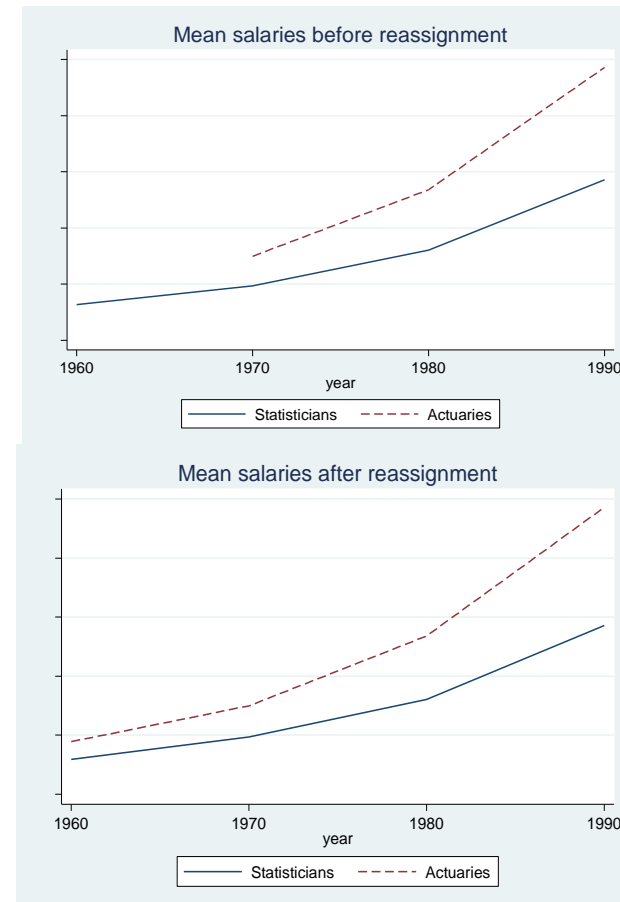
Imputing from 1970-1980 dual-coded data set

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

Desirable attributes of a classification

- Why work this arcane problem?
  - More accurate statistician category, by later definition
  - Longer time series for actuaries
  - Reduces SPARSENESS
    - One prefers a classification not be *sparse*, meaning it does not have many empty occ-year cells
  - Builds a technique
- One would like, for each occupation, accurate and well-behaved time series of:
  - mean wage
  - wage variance
  - fraction of the population

Nagging issues and next steps

- How biased will results be when built on occupations imputed this way?
  - ▶ I think: only a little
- What are other countries doing?
- Further exploit dual-coded 1970-80 and 1990-2000 data sets
  - ▶ How to systematize this better?
- Share improved *occ1990* variable (at IPUMS.org) and source code
- Start on 2010 classification as data comes in

Contact Information

Peter B. Meyer (Office of Productivity and Technology)
202-691-5678
meyer.peter@bls.gov

BLS
U.S. BUREAU OF LABOR STATISTICS
125 YEARS