Augmenting U.S. Census data on industry and occupation of respondents

Peter B. Meyer and Kendra Asher

- **U.S. Census Bureau classifies respondents in the Censuses of Population, and the Current Population Survey into industry and occupation categories**
- **The classifications change periodically, breaking time series**
- **Crosswalks and unified category systems bridge the periods, but combine too much and leave empty cells**
- **Here we predict industry and occupation classifications for respondents based on microdata from another period**
- **Some microdata available on employer type, job, income, location, age, sex, and race**
- **Some training data from surveys has been coded into two classification systems ("dual-coded data")**
- **We can set thresholds so imputed occupations or industries match various benchmarks from other data sources.**
- **We can test resulting "augmented" data sets on known trends, smoothness criteria, and population benchmarks.**
- **We can then adjust thresholds or focus study on particular categories.**
- **Thus we can add value to past Censuses**
- For IEEE DSAA conference, October 5-8, 2019. Findings are preliminary. Views are those of the authors only.

Occupation categories

Over time there are more categories and their content has changed. For example, electrical engineers grew as a category then split into subcategories. "Apprentices" were substantial categories which have disappeared. Many occupations are not mapped to exactly one in the next decade.

Occupational categories in the decennial Censuses, 1950-2010						
1950	**1960**	**1970**	**1980**	**1990**	**2000**	**2010**
287	297	441	503	501	509	539

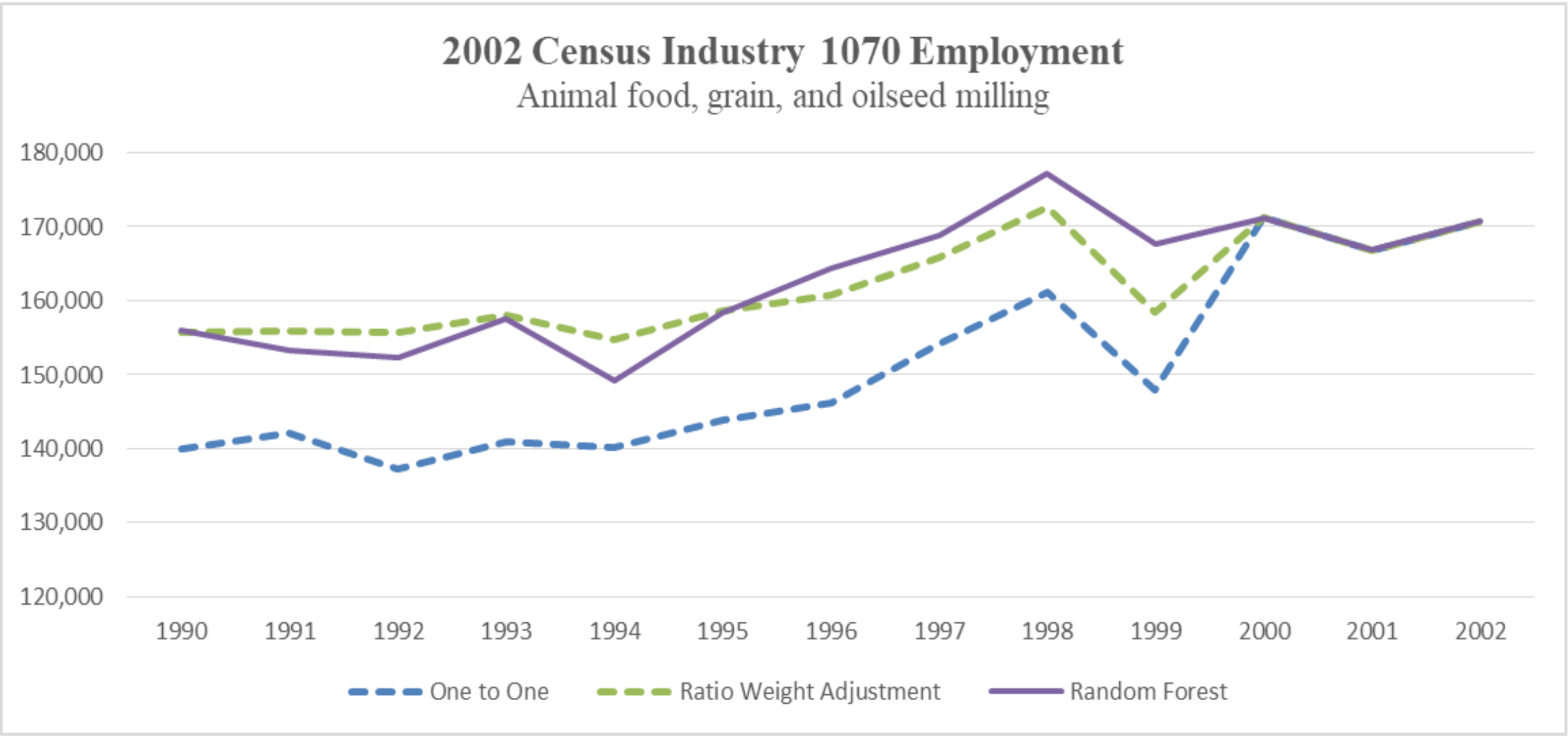
We have dual-coded data for the 1970-1980 transition and for the 1990-2000 transition. These give the "right answer" as determined by the definitive specialists at Census. Those training sets enable predictions in other data sets – here, from the Census of Population.

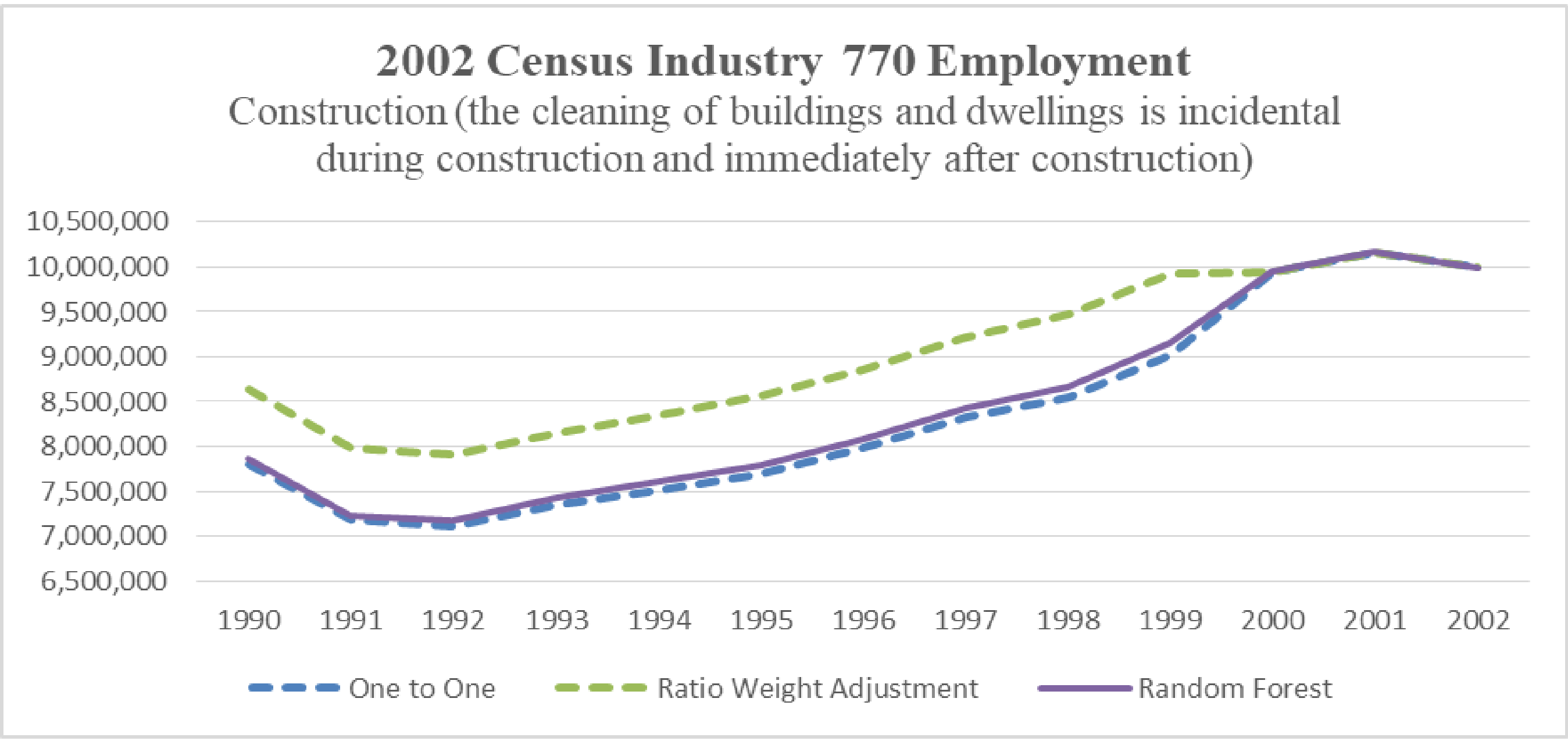
We examined several "source" occupations and predictors of how they would be classified in other decades. A logit regression gives an index of their likelihood to be in a destination category. A series of logits can handle more than two destination categories. Thresholds set by other evidence determine how many are assigned to each destination category. Examples:

Source category	Destination categories	Correlates / Predictors	Accuracy in test set
1960 Census "Statisticians and actuaries"	Statistician	Government employee	88% on 1970-1990 Censuses
	Actuary	In services industries; high business income; lives in state with big insurance co.	
1960 Census "Lawyers and judges"	Lawyer	Private sector; younger	84% on 1970-1990 Censuses
	Judge	State govt didn't finish college; low business income	
2000 Census "Appraisers and assessors of real estate"	1990: Real estate sales occupations	Self-employed or in "real estate" industry	90% in dual-coded 1990-2000 sample
	1990: Public officials	In "public finance" industry	
	1990: Managers and administrators, n.e.c.	In other professional or administrative industry	

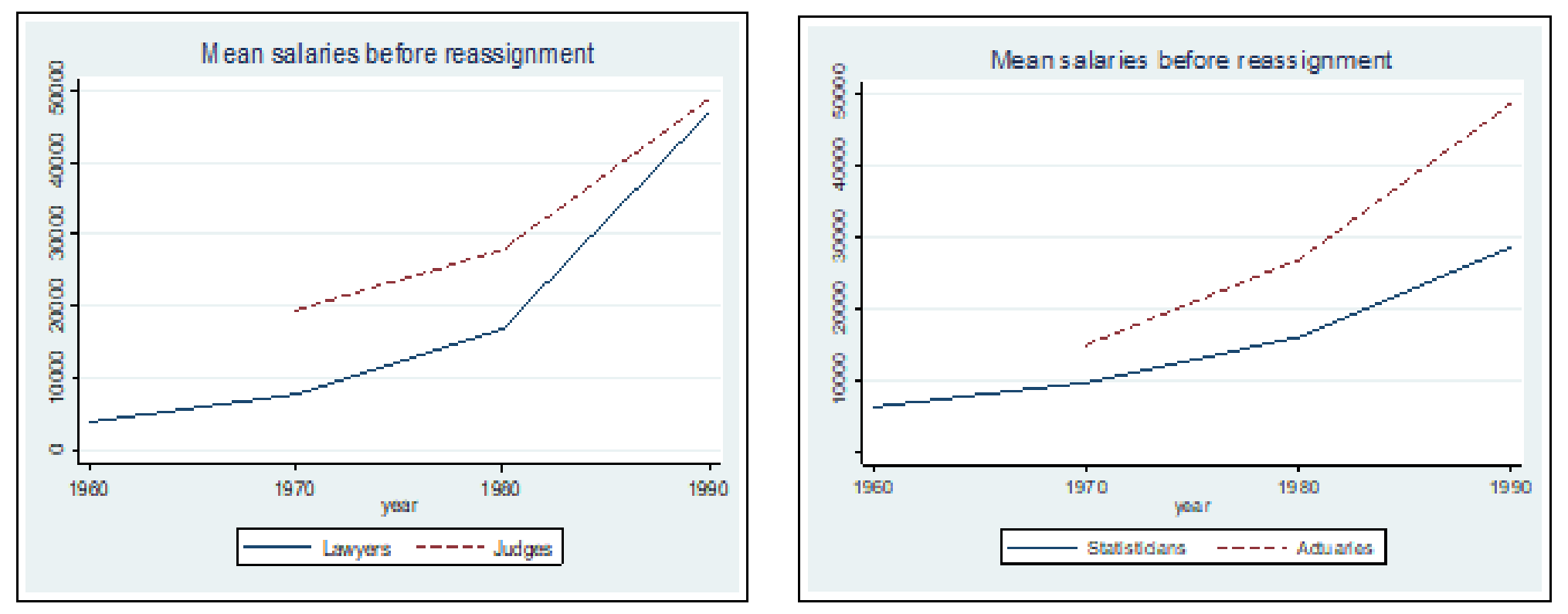
We intend to scale up to more occupations, using methods other than logit. Occupation and industry growth rates and earnings levels should be smooth over time. Predicting occupation and industry jointly can be more accurate. Economy-wide benchmarks will reduce biases toward common categories.

Estimates from industry data

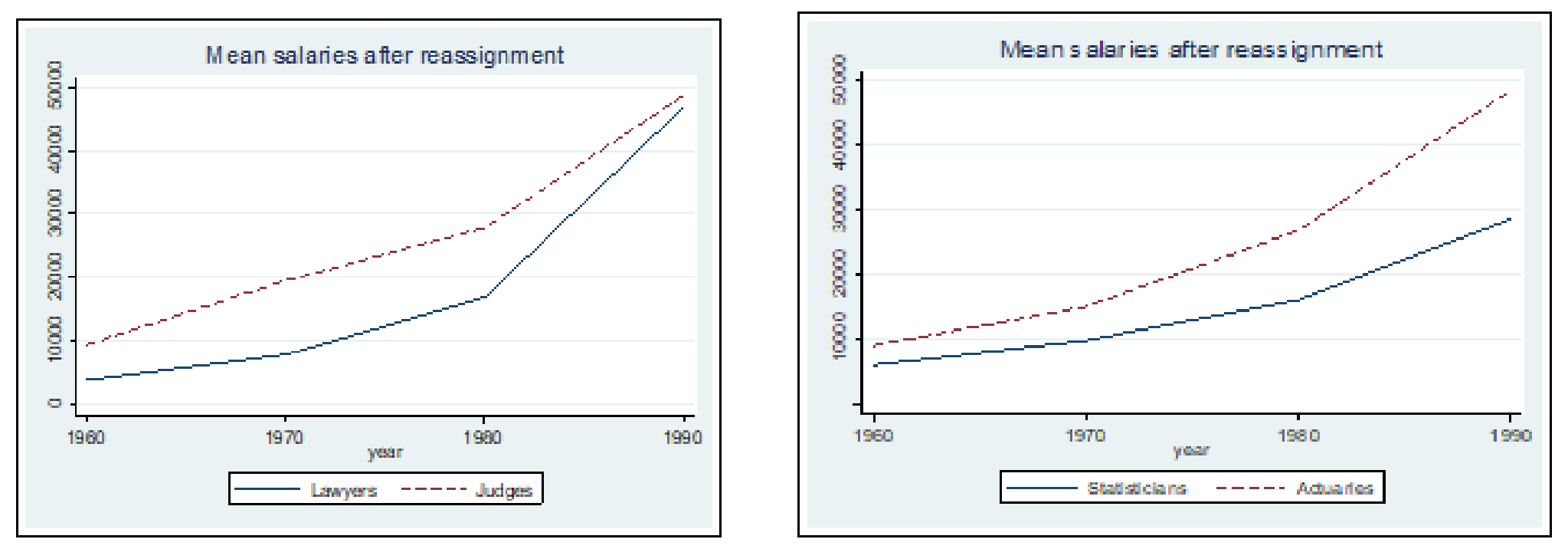




Estimates from occupation data



Earnings estimates from decennial Census data.



After imputations of judges and actuaries, we can fill in missing estimates and accuracy is slightly improved.

Industry classification breaks

Changes to Census industry classifications in 2000 and 2002 created a major break in Census and CPS data. Census industry classifications were converted to the 1997 North American Industrial Classification System (NAICS). It's possible to bridge data over that break with one-to-one industry crosswalks, or by taking ratios of totals from each source category and totaling them. Our predictive assignments of the destination industry for *each* underlying observation will be more precise and accurate.

The one-to-one method puts all the observations from each 1990 industry into the same 2002 industry. But most 2002 industries are not composed of a single 1990 industry, and most 1990 industry observations convert into more than one 2002 industry. For example, we find in our dual-coded data set that 70 different 1990 Census industries contribute 1% or more of their observations' weight to the 2002 Census industry 770, for most of "construction." Of these, 1990 Census industry 60, also titled "construction," is the largest contributor.

Another method of conversion is to create ratios based on the sum of weights of the dual-coded CPS data set, split the observations classified in 1990 Census industries, and adjust the split observations' weights based on the created ratios. This method, "ratio weight adjustment," is more accurate than the one-to-one method, but it induces error by drawing from observations that don't belong in the destination industry.

In the dual-coded data set, 2002 Census industry 1070, titled "animal food, grain, and oilseed mill," contains observations from six 1990 Census industries at 1% or more of their weight, listed below.

1990 industries contributing 1% of more to 2002 Census industry 1070		
1990 industry	**1990 Title**	**Ratio weight adjustment**
102	Canned, frozen, and preserved fruits and vegetables	1.56
110	Grain mill products	86.87
121	Misc. food preparations and kindred products	10.41
122	Not specified food industries	5.79
551	Farm-product raw materials	2.20
561	Farm supplies	4.67

Random forest algorithms improve on this, using 500-1,000 decision trees to predict the 2002 Census industry for each individual in the 1990s data. Our random forest algorithms use 17-20 features of the data, notably occupation, demographics, and location. Accuracy seems better for larger industries.

BUREAU OF LABOR STATISTICS | OFFICE OF PRODUCTIVITY AND TECHNOLOGY | productivity@bls.gov | @BLS_gov