A Framework for Sharing Confidential Research Data, Applied to Investigating Differential Pay by Race in the U. S. Government

Supplement: Synthetic Data Validation

Duke University Synthetic Data Project

March 3, 2018

The following graphs and tables are excerpted from work done by the Synthetic Data Team at Duke University to validate the DIBBS synthetic federal employee data set with corresponding authentic data supplied by the U.S. Office of Personnel Management (OPM).[1] The selection here highlights two and three level covariate relationships, especially involving important research variables such as sex, race, age, education, agency, occupation, year, and pay. In assessing similarity of the data sets, emphasis is placed on utility, or the degree to which answers to meaningful research questions obtained from use of synthetic data agree with those from use of corresponding authentic data. Graphs and tables representing synthetic data contain the text "DIBBS" while those for authentic data contain either "OPM" or "JdF."[2] All codes and definitions are taken from the U.S. Office of Personnel Management Guide to Data Standards (U.S. Office of Personnel Management, GDS). For additional information and guidance on use and interpretation of data made available by OPM, see (U.S. Office of Personnel Management, Data Analysis).

TWO VARIABLE CORRELATION

Figure 1 shows, for pay plan GS, full-time observations, correlations between 1.) the variable indicated in the title bars and 2.) all levels of all other variables in title bars. Synthetic variable pair correlations are plotted (y-axis) against corresponding pair correlations in the authentic data (x-axis). Points lying near the reference line (slope of 1.0) indicate equality between data sets. Agency and occupation are truncated to the first two positions. Note that correlations involving categorical variables, or fixed effects, effectively measure the association of proportion of observations with levels of the second variable. Missing counts are the number of variable level combinations that appear in the other data set but not in the one indicating a count. For instance, JdF=2 in the agency panel would indicate that observations exist in the synthetic data, but not in the authentic data, for two combinations of agency and some level of a second variable. Note that all missing counts are a multiple of three. This is due to agencies AL, CP, and GD missing in the synthetic data.

Observation: Correlation of pairs of levels of variables within synthetic data are very near those of corresponding pairs in the authentic data. This is indicated by the near proximity of all plotted points to the reference line of slope 1.0, including those for extreme correlation values.
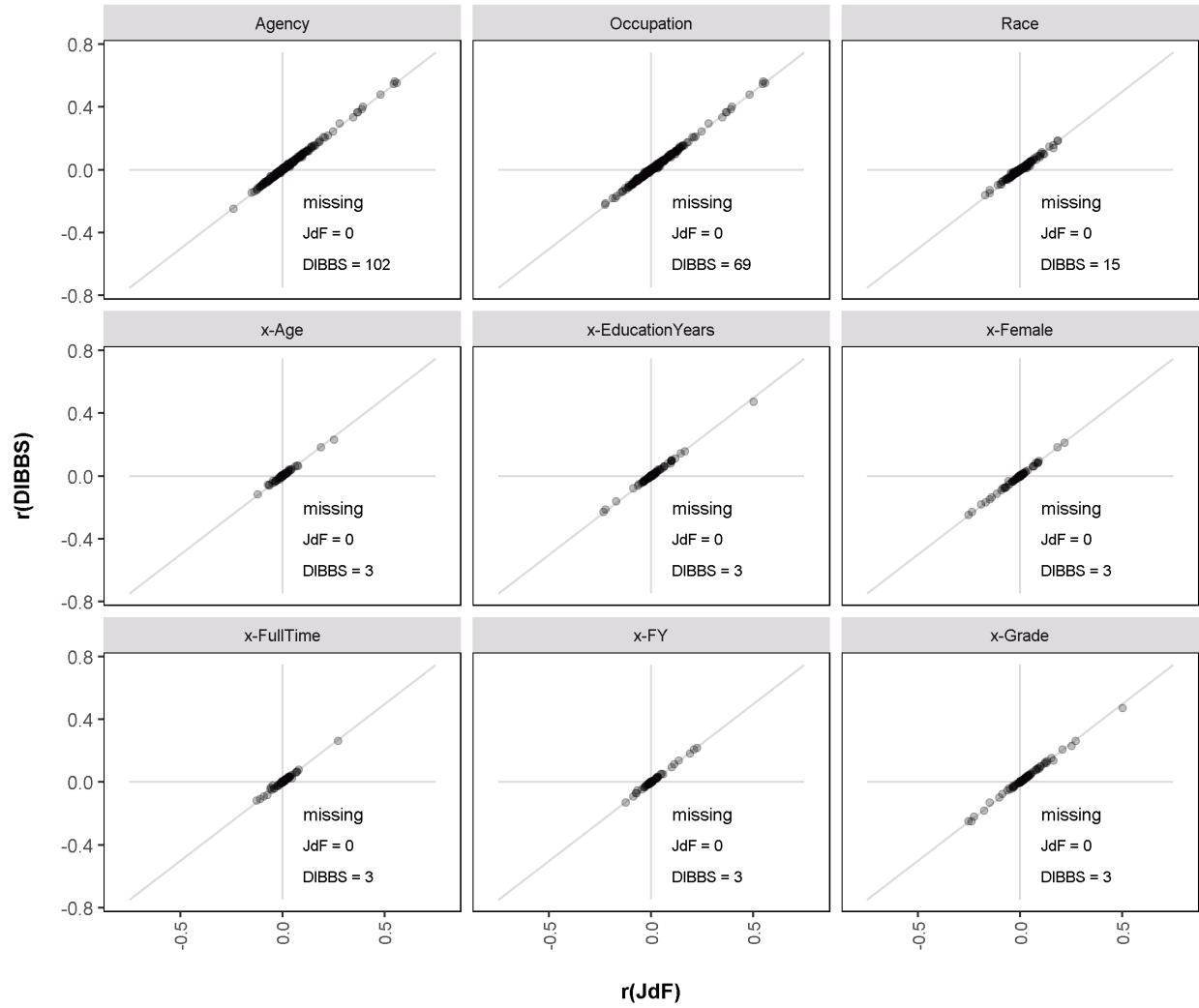

Figure 1: Two variable correlations of corresponding levels of synthetic and authentic data. Synthetic level correlation on y-axis, corresponding authentic correlation on x-axis.

CORRELATION OF PRIMARY VARIABLES WITH TWO-VARIABLE INTERACTIONS

Figures 2 and 3 show, for pay plan GS, full-time observations, correlations between 1.) the variable indicated in the graph title, 2.) all combinations of levels of the variable listed in a title bar, and 3.) all levels of other variables appearing in the title bars. These constitute correlation of main variables with two variable interactions. In the case of categorical variables, or fixed effects, this is the association of a primary variable with the proportion of observations in interacting level combinations of two other variables. Agency and occupation truncated to first two positions.

Observation: Proximity of all points to slope 1.0 reference line indicates agreement of three-variable associations between data sets and implies depth of utility beyond simple pairwise relationships.

(a) Correlations involving sex

(b) Correlations involving race

(c) Correlations involving agency
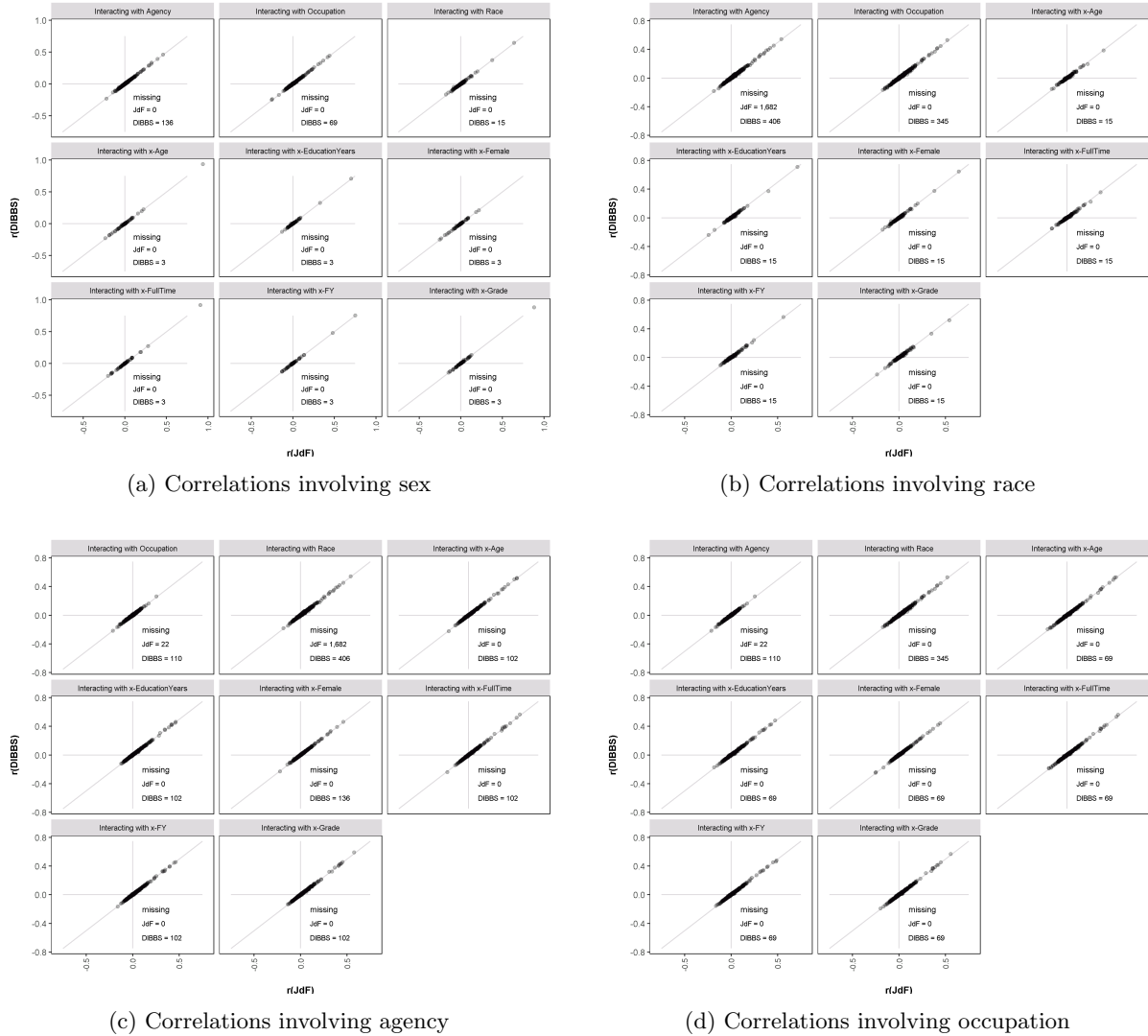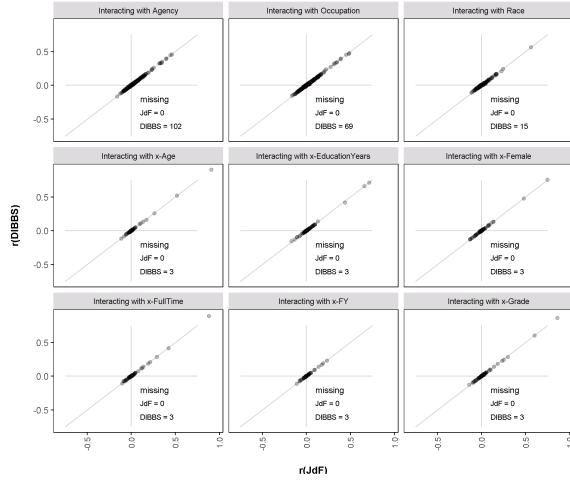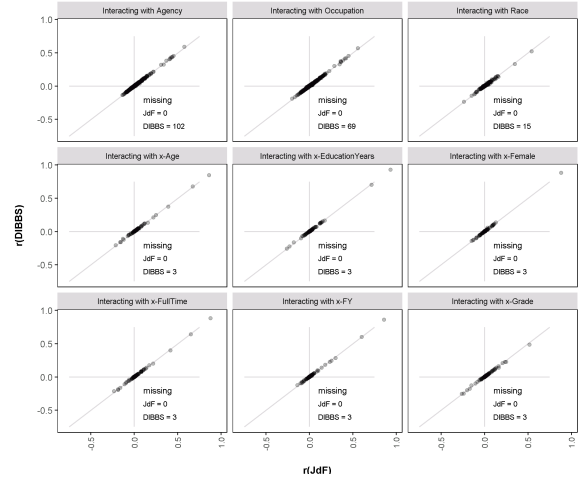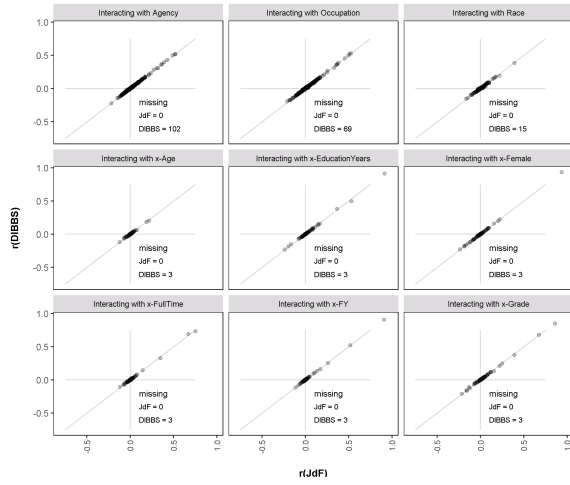
(d) Correlations involving occupation

Figure 2: Correlations of primary variables with two variable interactions. Variable set one. Synthetic level correlation on y-axis, corresponding authentic correlation on x-axis.
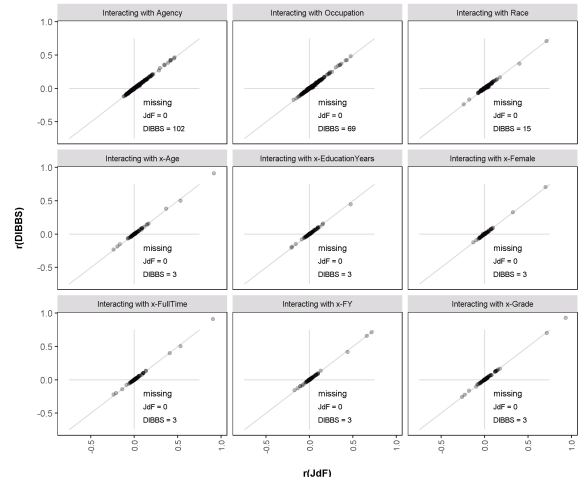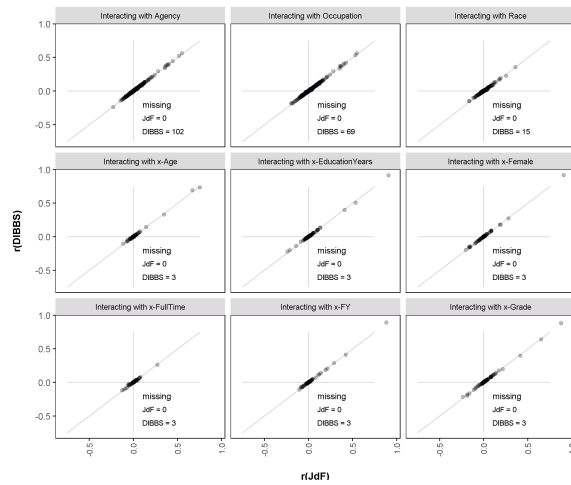
(a) Correlations involving Fiscal Year

(b) Correlations involving Grade

(c) Correlations involving Age

(d) Correlations involving Education

(e) Correlations involving Work Schedule

Figure 3: Correlations of primary variables with two variable interactions. Variable set two. Synthetic level correlation on y-axis, corresponding authentic correlation on x-axis.

CUMULATIVE MASS (PROPORTION OBSERVATIONS) BY PAY PLAN AND OCCUPATION

Figures 4 and 5 contain example CMF plots of pay plan and occupation combinations. All occupations within each pay plan are represented. Solid line for authentic data, dashed line for synthetic. Overlapping or nearness of lines indicates equality of cumulative mass for corresponding levels of occupation within pay plan. "nJ" indicates observation count in authentic data, "nD" indicates synthetic data observation count. Near identical distribution is observed for high frequency pay plans GS, WG, GM, and VN, which account for more than 95% of observations, indicating overall close agreement between data sets. Increasing departure observed as number of observations decreases.
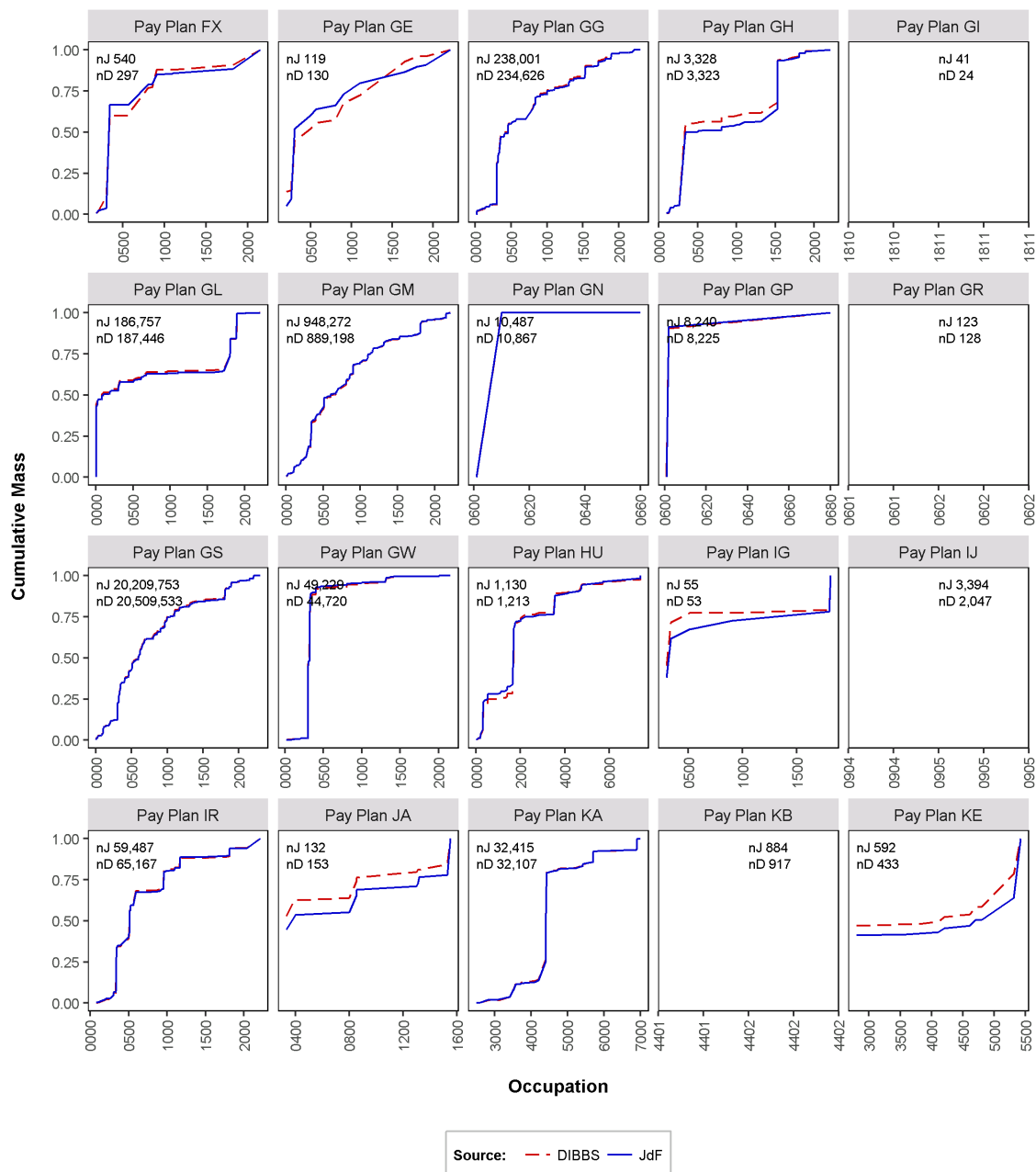

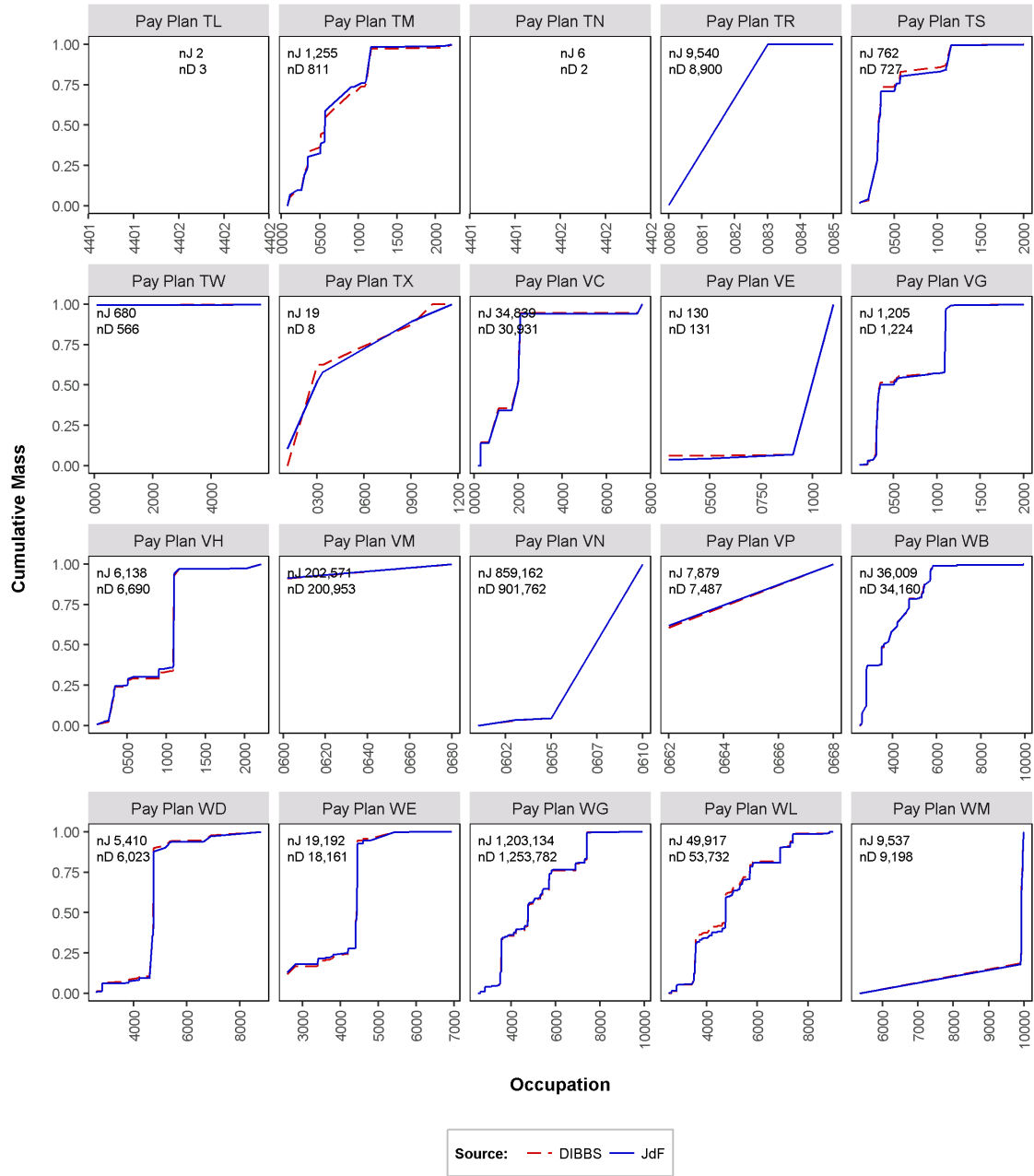Figure 4: Cumulative mass by occupation within pay plan. Pay plan set one.

Figure 5: Cumulative mass by occupation within pay plan. Pay plan set 2.

DISTRIBUTION OF BASIC PAY BY AGENCY

Basic pay is an important dependent study variable and the distribution of pay values in the authentic data must be maintained in the synthetic data. Figure 6 plots the distribution of basic pay for the top eight frequency agencies (first two positions): Department of Agriculture (AG), Department of Justice (DJ), Department of Health and Human Services (HE), Department of Homeland Security (HS), Department of Interior (IN), Department of Transportation (TD), Department of Treasury (TR), and the Department of Veterans Affairs (VA). These agencies account for approximately 85% of observations. Synthetic distribution represented by dashed line, authentic distribution by solid line. "n(D)" indicates synthetic data observation frequency, "n(J)" indicates authentic data observation count.

Observations: Although each data set is represented in each graph, a single striped-appearing line is visible, due to identical frequency proportions at each pay level. Local increases, decreases, and trends in authentic distribution are accurately represented in the synthetic data.
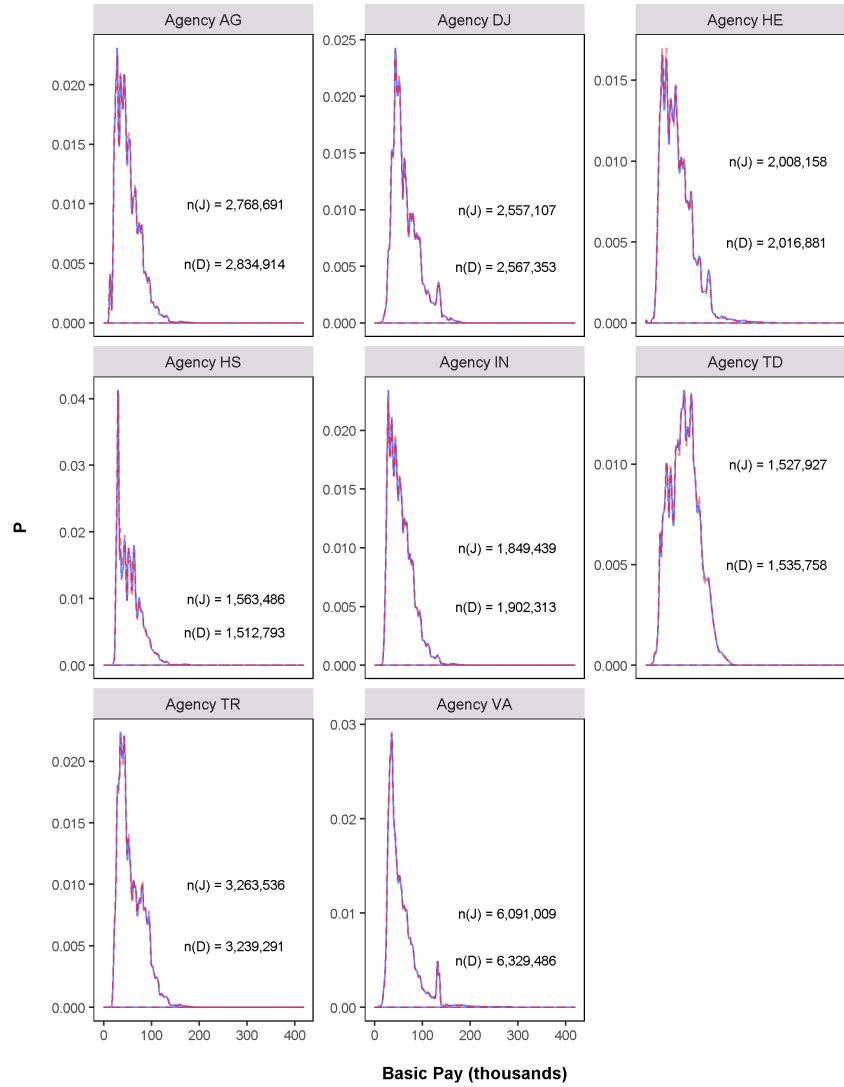

Figure 6: Basic pay marginal distribution for top eight agencies. Dashed line for synthetic data, solid line for authentic.

DISTRIBUTION OF BASIC PAY BY PROFESSIONAL, SUPERVISORY, COLLEGE EDUCATION, AND
WORK SCHEDULE CATEGORY

Professional classification, supervisory status, and college education are important independent variables in
human capital research. Figures 7 through 14 plot, for the top eight frequency agencies, the distribution of
basic pay by these independent variables and work schedule code. One column for each professional, supervi-
sory, college combination (column code position one equals "P" if occupational category is administrative or
professional, position two equals "S" if supervisory status is enabled, position three equals "C" if education
level at or above college). One row for each work schedule code [significant codes are full time (F), full time
seasonal (G), intermittent (I), intermittent seasonal (J), and part time (P)]. Synthetic distribution indicated
by dashed line, authentic distribution by solid line. "n(D)" indicates synthetic data observation frequency,
"n(J)" indicates authentic data observation count.

Observations: There exists near identical distribution for high frequency combinations, as indicated by
striped, single line appearance due to overlay of synthetic on authentic lines. Slight differences in distribution
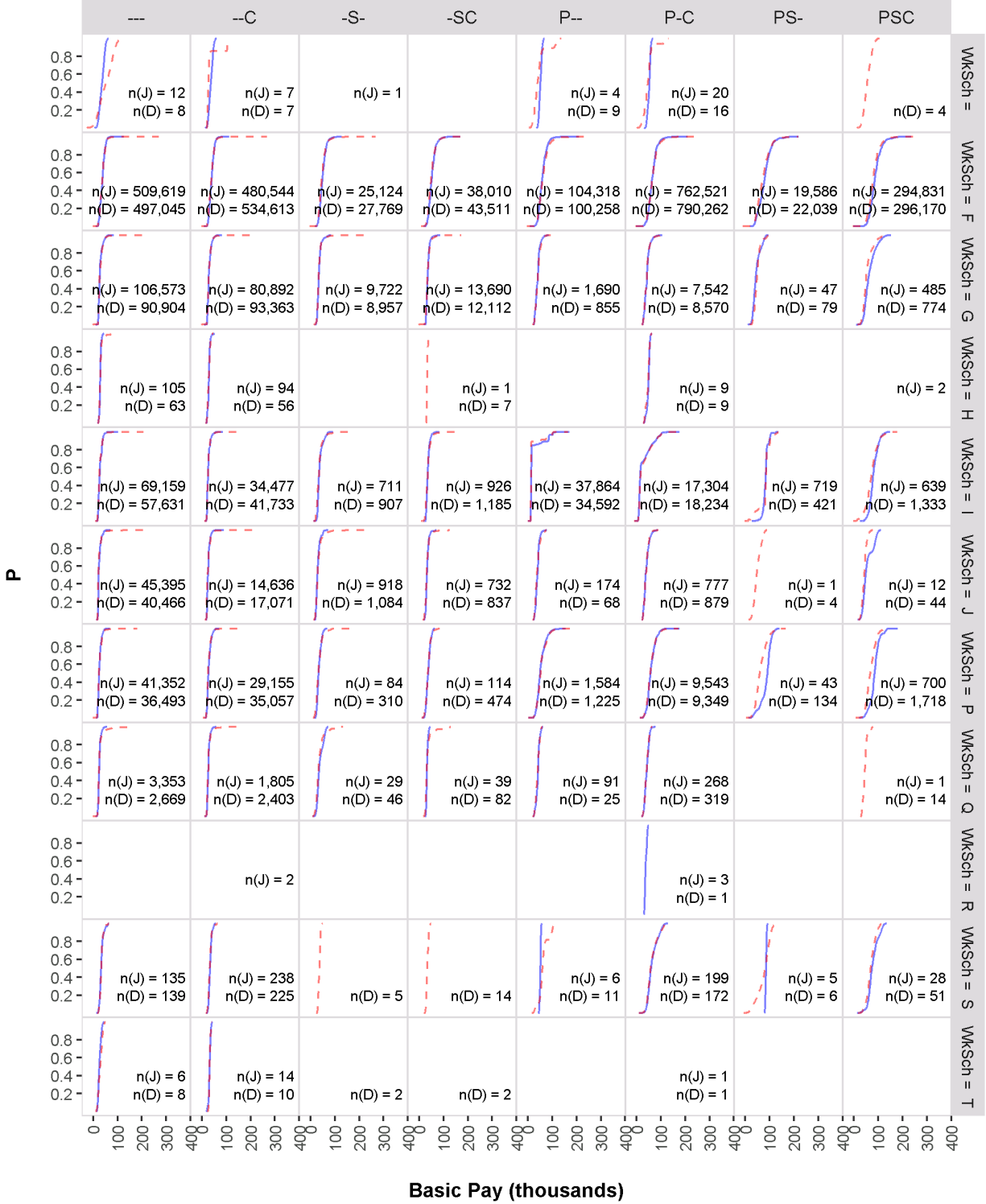are observed for small frequency combinations.

Figure 7: Basic pay distribution by joint professional, supervisory status, college education, and work schedule categories. Department of Agriculture (AG). Dashed line for synthetic data, solid line for authentic.
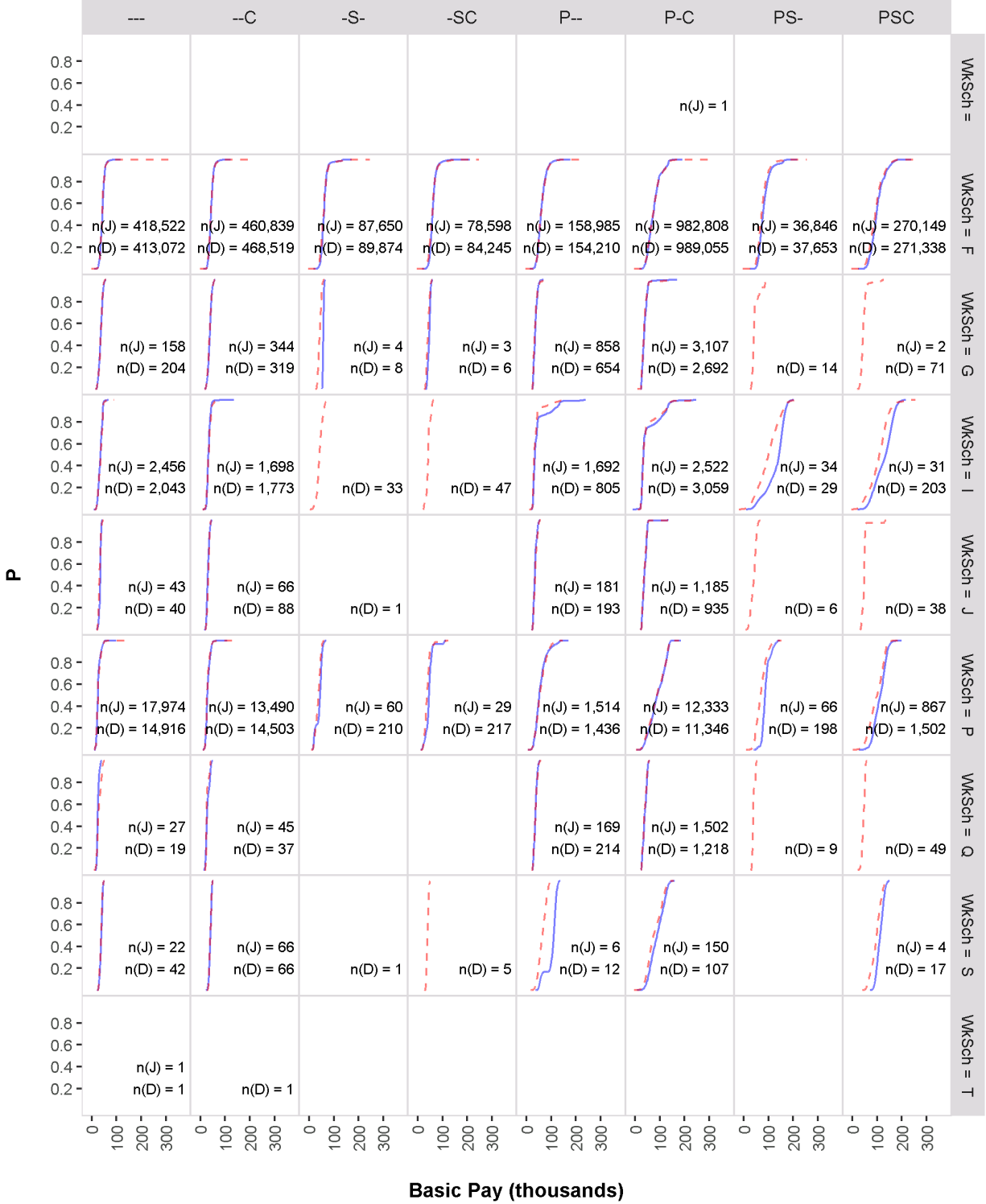
Figure 8: Basic pay distribution by joint professional, supervisory status, college education, and work schedule categories. Department of Justice (DJ). Dashed line for synthetic data, solid line for authentic.
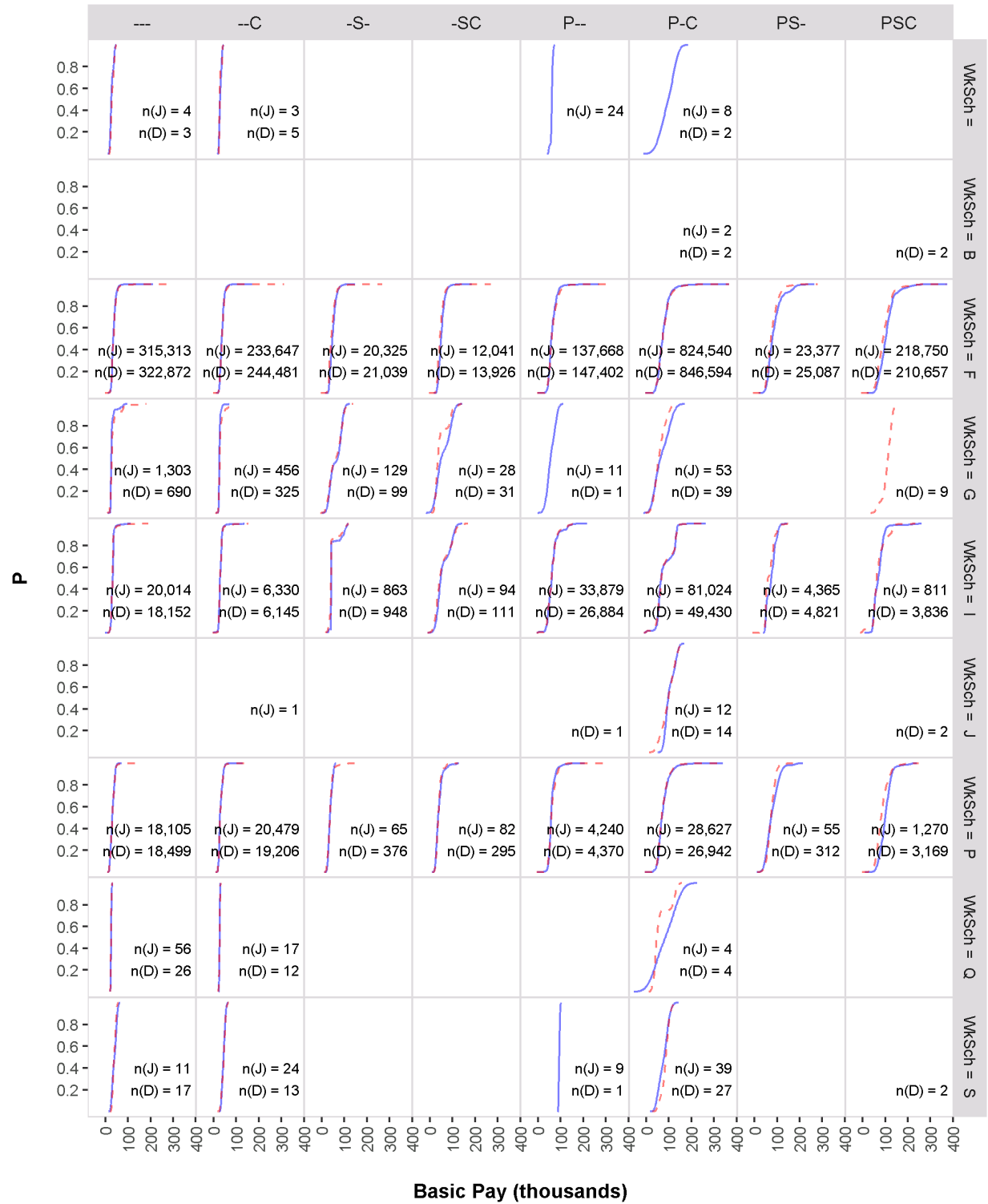
Figure 9: Basic pay distribution by joint professional, supervisory status, college education, and work schedule categories. Department of Health and Human Services (HE). Dashed line for synthetic data, solid line for authentic.
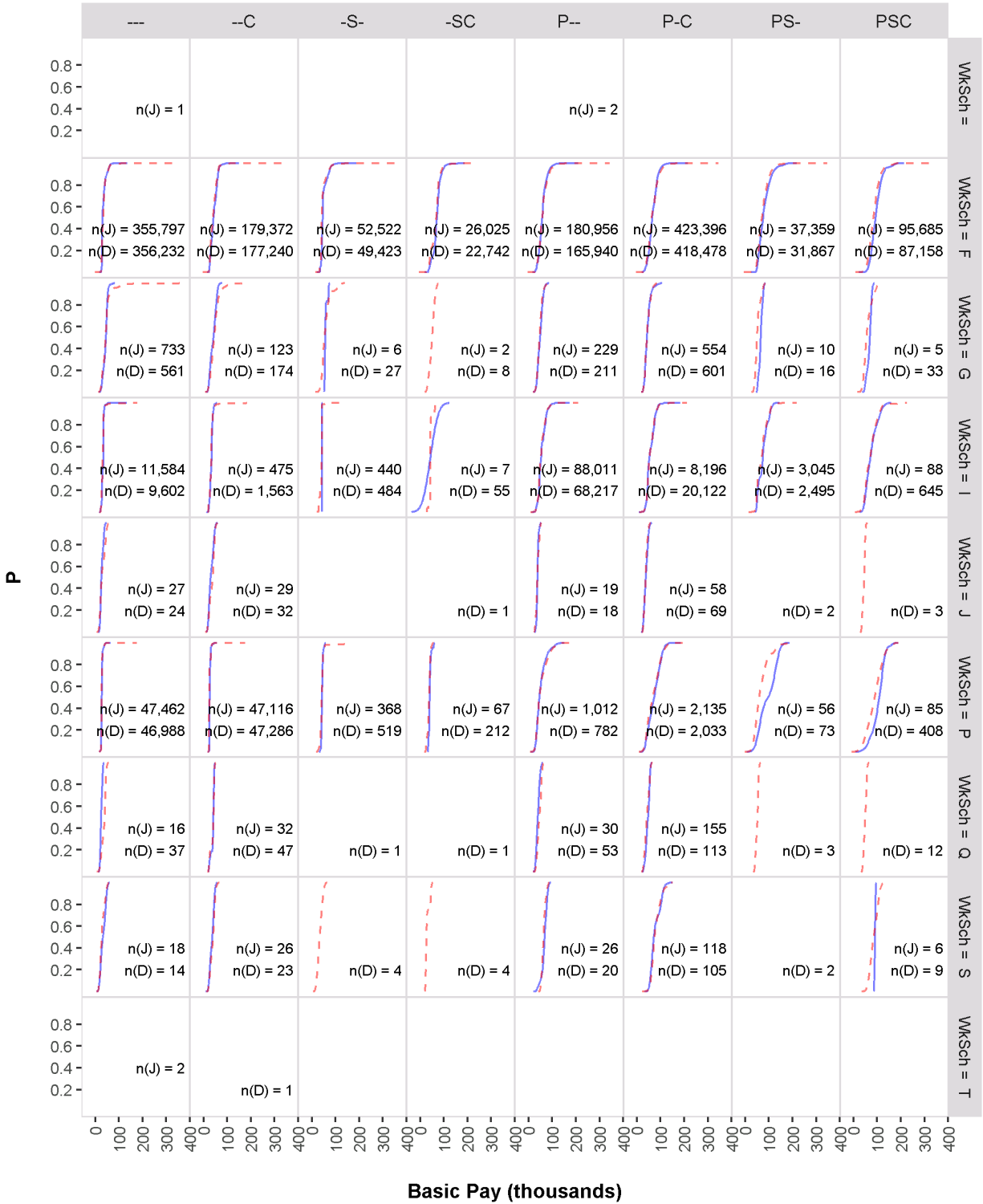
	---	--C	-S-	-SC	P--	P-C	PS-	PSC

WkSch =

n(J) = 1 (---)   n(J) = 2 (P--)

WkSch = F

| n(J) = 355,797 | n(J) = 179,372 | n(J) = 52,522 | n(J) = 26,025 | n(J) = 180,956 | n(J) = 423,396 | n(J) = 37,359 | n(J) = 95,685 |
| n(D) = 356,232 | n(D) = 177,240 | n(D) = 49,423 | n(D) = 22,742 | n(D) = 165,940 | n(D) = 418,478 | n(D) = 31,867 | n(D) = 87,158 |

WkSch = G

| n(J) = 733 | n(J) = 123 | n(J) = 6 | n(J) = 2 | n(J) = 229 | n(J) = 554 | n(J) = 10 | n(J) = 5 |
| n(D) = 561 | n(D) = 174 | n(D) = 27 | n(D) = 8 | n(D) = 211 | n(D) = 601 | n(D) = 16 | n(D) = 33 |

WkSch = I

| n(J) = 11,584 | n(J) = 475 | n(J) = 440 | n(J) = 7 | n(J) = 88,011 | n(J) = 8,196 | n(J) = 3,045 | n(J) = 88 |
| n(D) = 9,602 | n(D) = 1,563 | n(D) = 484 | n(D) = 55 | n(D) = 68,217 | n(D) = 20,122 | n(D) = 2,495 | n(D) = 645 |

WkSch = J

| n(J) = 27 | n(J) = 29 | | | n(J) = 19 | n(J) = 58 | | |
| n(D) = 24 | n(D) = 32 | | n(D) = 1 | n(D) = 18 | n(D) = 69 | n(D) = 2 | n(D) = 3 |

WkSch = P

| n(J) = 47,462 | n(J) = 47,116 | n(J) = 368 | n(J) = 67 | n(J) = 1,012 | n(J) = 2,135 | n(J) = 56 | n(J) = 85 |
| n(D) = 46,988 | n(D) = 47,286 | n(D) = 519 | n(D) = 212 | n(D) = 782 | n(D) = 2,033 | n(D) = 73 | n(D) = 408 |

WkSch = Q

| n(J) = 16 | n(J) = 32 | | | n(J) = 30 | n(J) = 155 | | |
| n(D) = 37 | n(D) = 47 | n(D) = 1 | n(D) = 1 | n(D) = 53 | n(D) = 113 | n(D) = 3 | n(D) = 12 |

WkSch = S

| n(J) = 18 | n(J) = 26 | | | n(J) = 26 | n(J) = 118 | | n(J) = 6 |
| n(D) = 14 | n(D) = 23 | n(D) = 4 | n(D) = 4 | n(D) = 20 | n(D) = 105 | n(D) = 2 | n(D) = 9 |

WkSch = T

| n(J) = 2 | n(D) = 1 | | | | | | |

Basic Pay (thousands)

Figure 10: Basic pay distribution by joint professional, supervisory status, college education, and work schedule categories. Department of Homeland Security (HS). Dashed line for synthetic data, solid line for authentic.
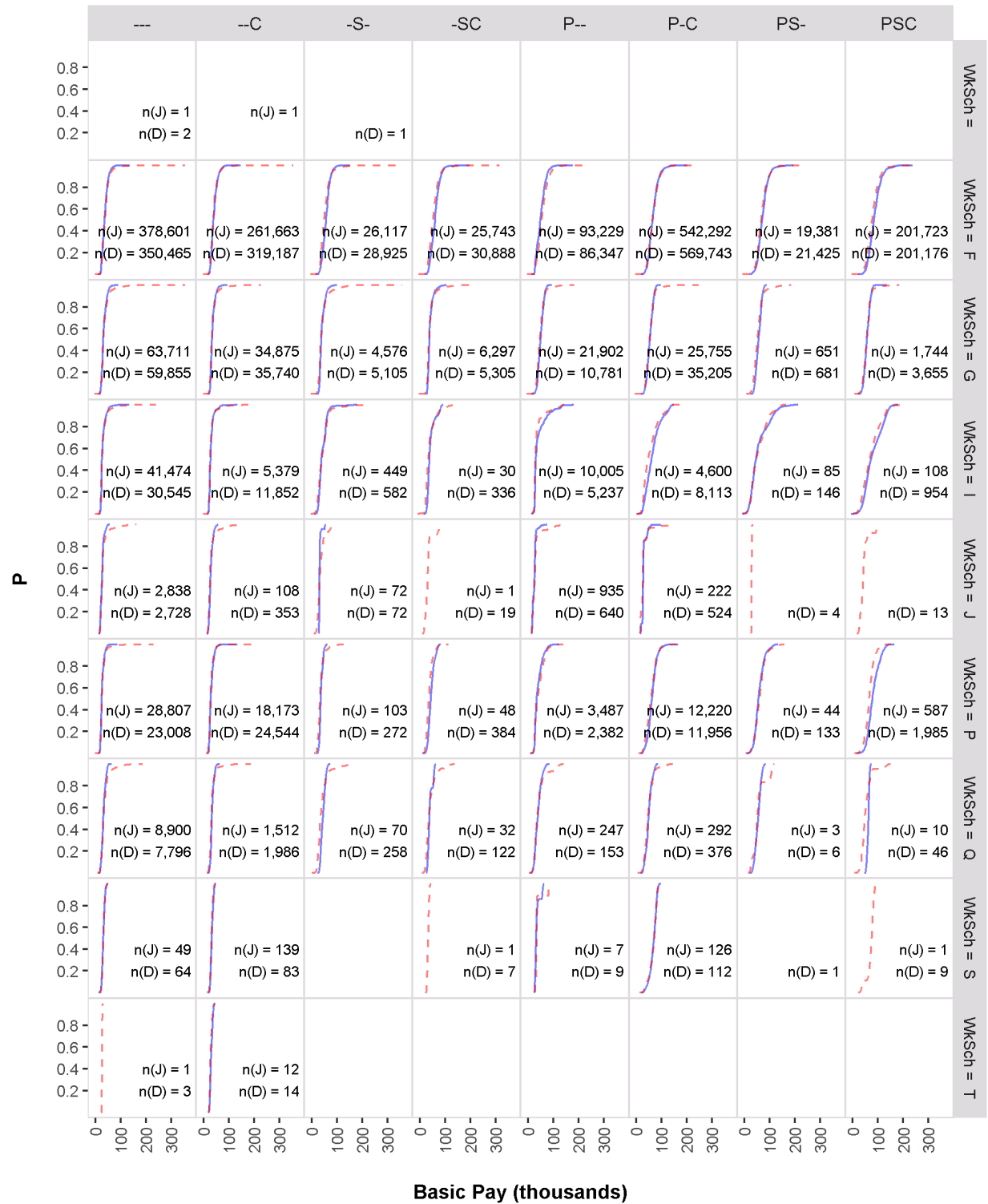
Figure 11: Basic pay distribution by joint professional, supervisory status, college education, and work schedule categories. Department of Interior (IN). Dashed line for synthetic data, solid line for authentic.
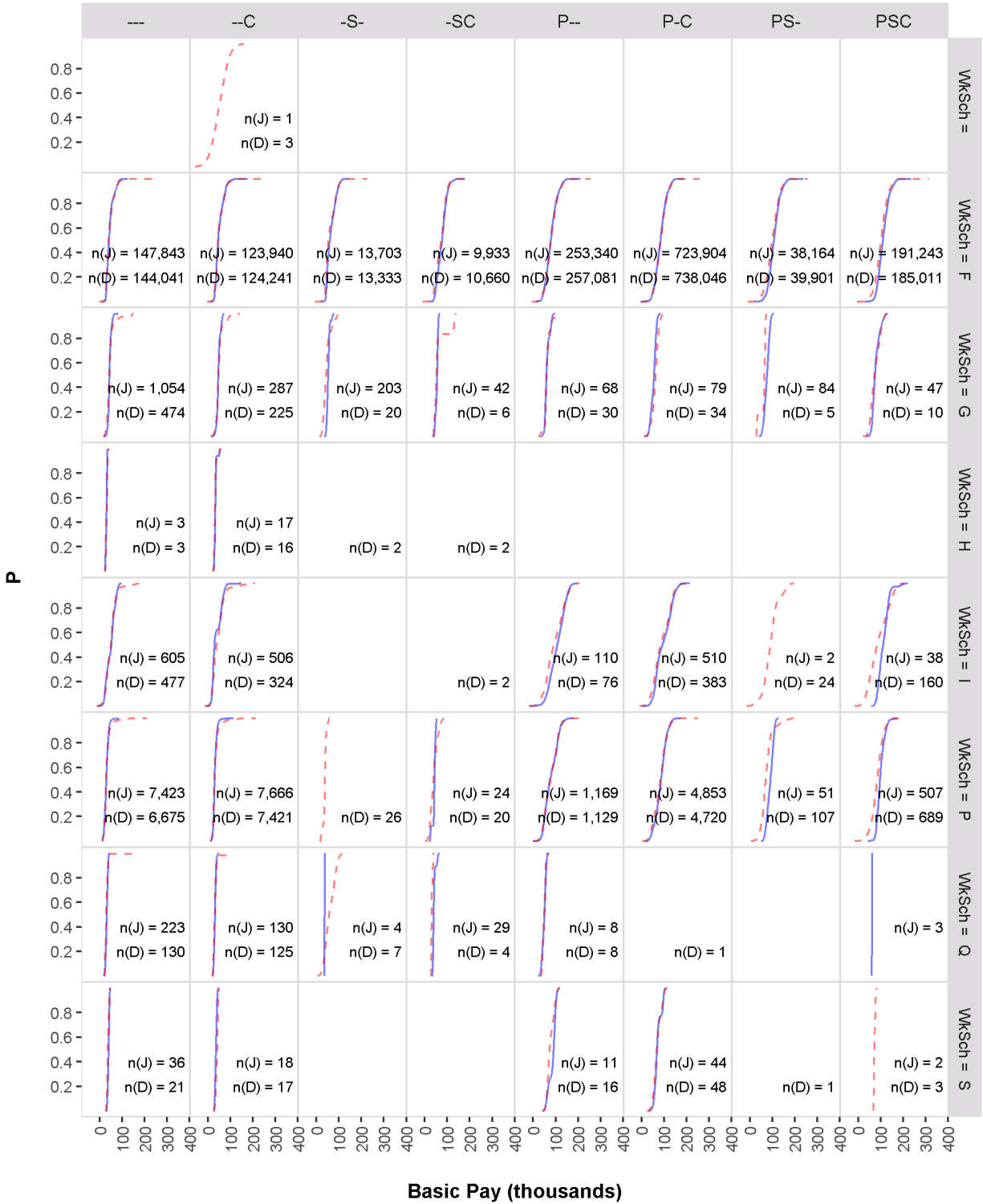
Figure 12: Basic pay distribution by joint professional, supervisory status, college education, and work schedule categories. Department of Transportation (TD). Dashed line for synthetic data, solid line for authentic.
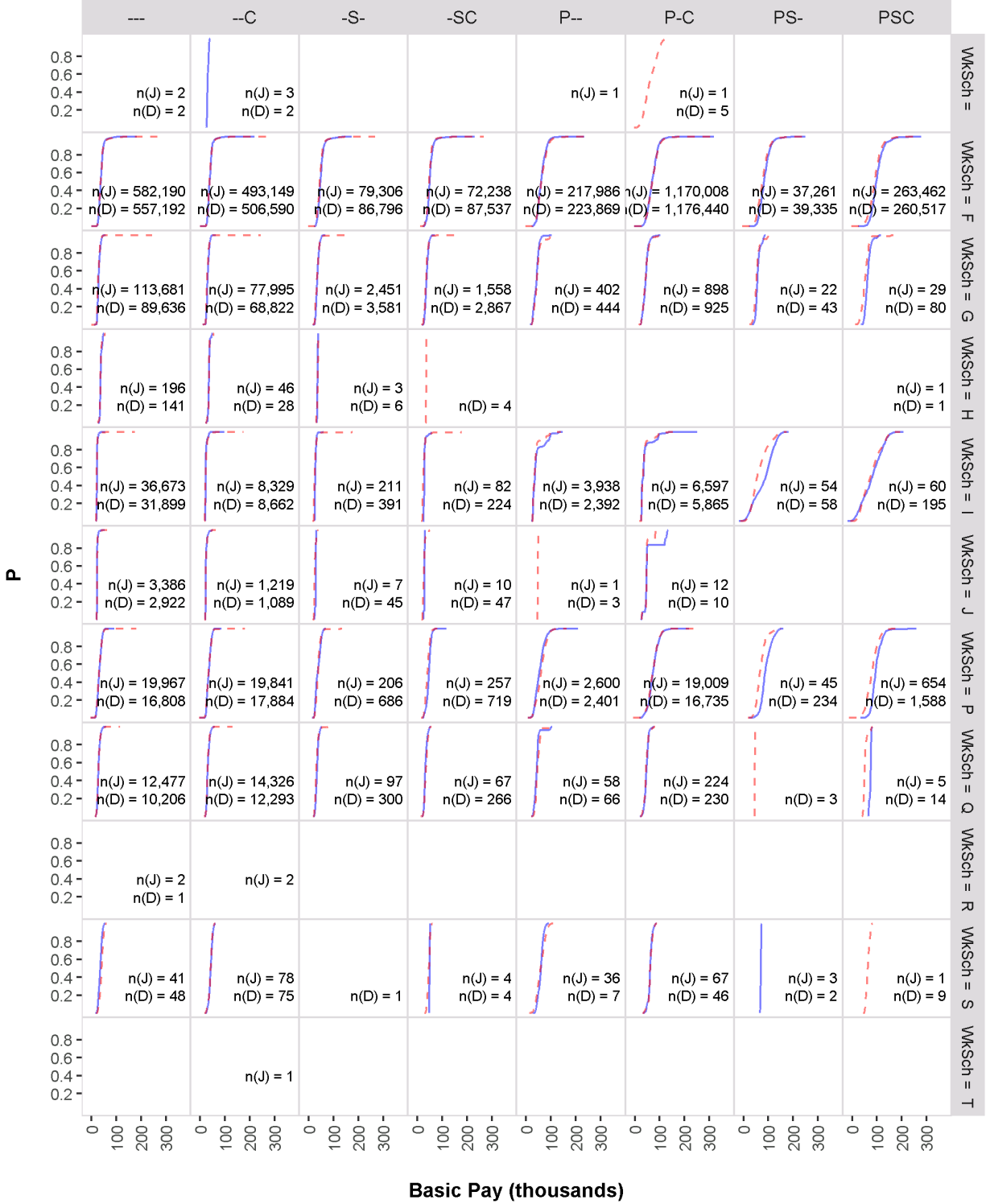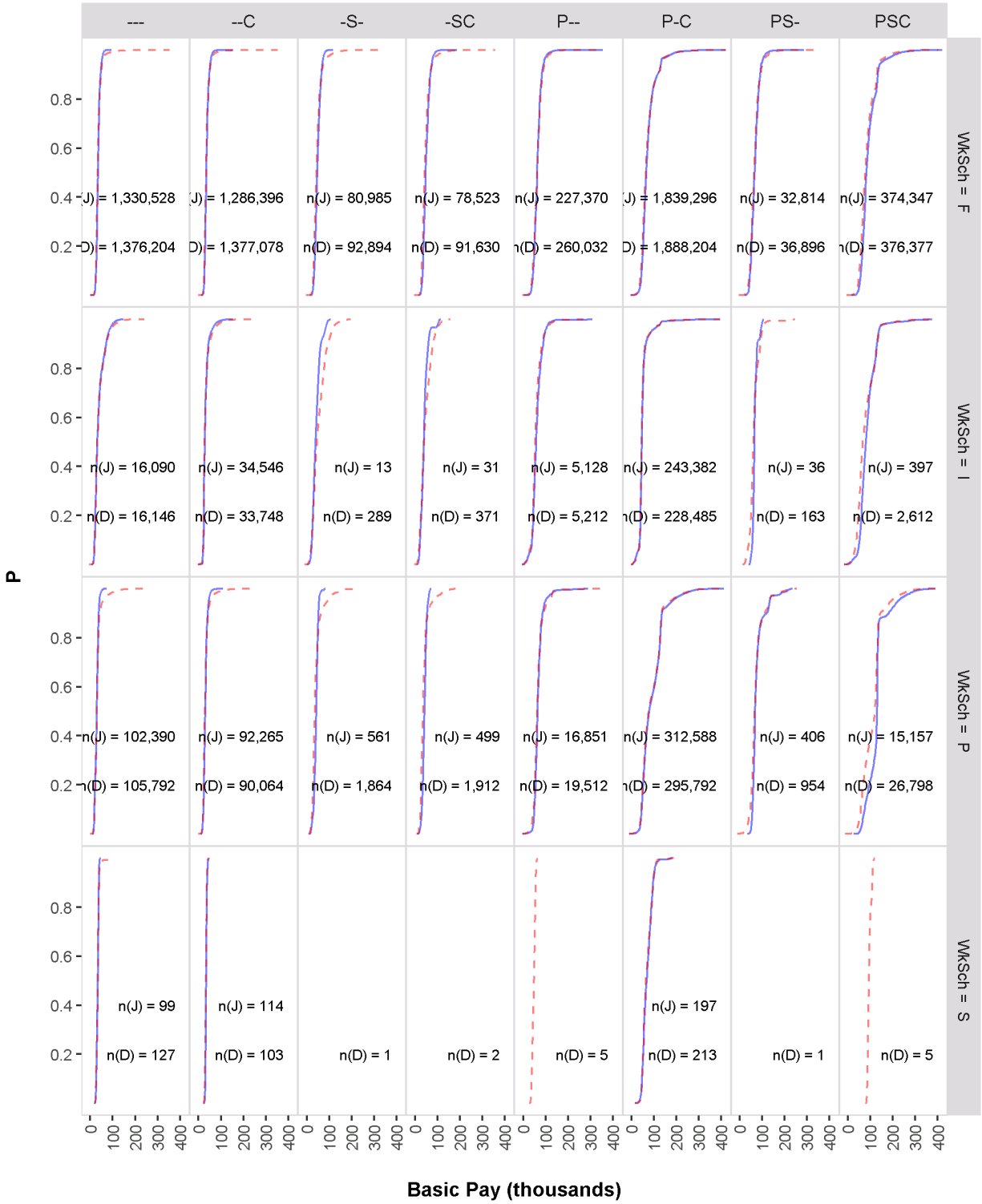
14

Figure 13: Basic pay distribution by joint professional, supervisory status, college education, and work schedule categories. Department of Treasury (TR). Dashed line for synthetic data, solid line for authentic.

	---	--C	-S-	-SC	P--	P-C	PS-	PSC
WkSch = F	n(J) = 1,330,528 n(D) = 1,376,204	n(J) = 1,286,396 n(D) = 1,377,078	n(J) = 80,985 n(D) = 92,894	n(J) = 78,523 n(D) = 91,630	n(J) = 227,370 n(D) = 260,032	n(J) = 1,839,296 n(D) = 1,888,204	n(J) = 32,814 n(D) = 36,896	n(J) = 374,347 n(D) = 376,377
WkSch = I	n(J) = 16,090 n(D) = 16,146	n(J) = 34,546 n(D) = 33,748	n(J) = 13 n(D) = 289	n(J) = 31 n(D) = 371	n(J) = 5,128 n(D) = 5,212	n(J) = 243,382 n(D) = 228,485	n(J) = 36 n(D) = 163	n(J) = 397 n(D) = 2,612
WkSch = P	n(J) = 102,390 n(D) = 105,792	n(J) = 92,265 n(D) = 90,064	n(J) = 561 n(D) = 1,864	n(J) = 499 n(D) = 1,912	n(J) = 16,851 n(D) = 19,512	n(J) = 312,588 n(D) = 295,792	n(J) = 406 n(D) = 954	n(J) = 15,157 n(D) = 26,798
WkSch = S	n(J) = 99 n(D) = 127	n(J) = 114 n(D) = 103	n(D) = 1	n(D) = 2	n(D) = 5	n(J) = 197 n(D) = 213	n(D) = 1	n(D) = 5

Figure 14: Basic pay distribution by joint professional, supervisory status, college education, and work schedule categories. Department of Veterans Affairs (VA). Dashed line for synthetic data, solid line for authentic.

DISTRIBUTION OF BASIC PAY BY OCCUPATION AND SUPERVISORY STATUS

803 distinct occupations are represented in the data supplied by OPM. To verify distribution of basic pay by occupation and supervisory status in the synthetic data, box plots consisting of a pair of authentic/synthetic distributions for each occupation were constructed. Figures 15 through 17 show box plots for the first 120 occupations in order of occupation code. Trade, or blue collar, occupations begin at code 2500. Figures 18 and 19 show the distributions of the first 80 of these occupations. Remaining occupations exhibit patterns similar to those presented.

Observations: Median pay and inter-quartile ranges appear consistent between data sets. Upper tails of distributions of in the synthetic data generally appear greater than corresponding authentic distributions, particulary for trade occupations. Note that, of the 318 trade occupations represented in the authentic data, 190 have proportion female observations less than 0.05. Given disparity in federal employee pay by gender (Bolton and de Figueiredo, 2017) and a requirement, for protection of privacy, of a degree of modification of gender and occupation in synthetic observations, differences in pay extremes may be expected.
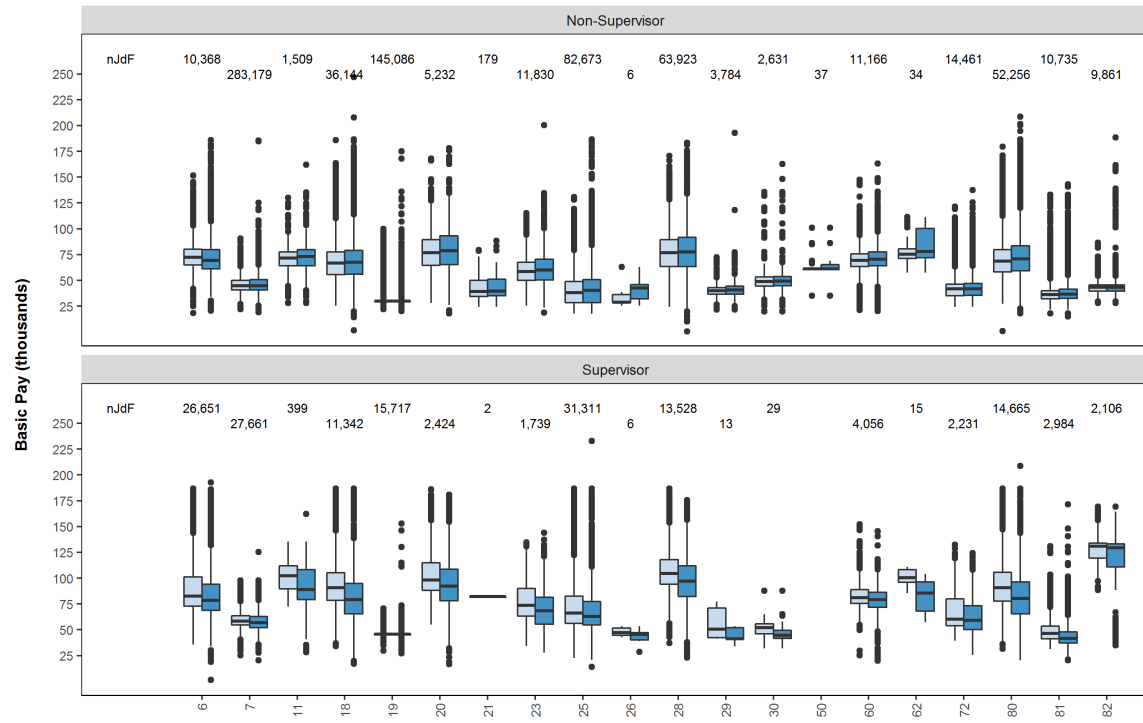
(a) Occupations 0006 through 0082
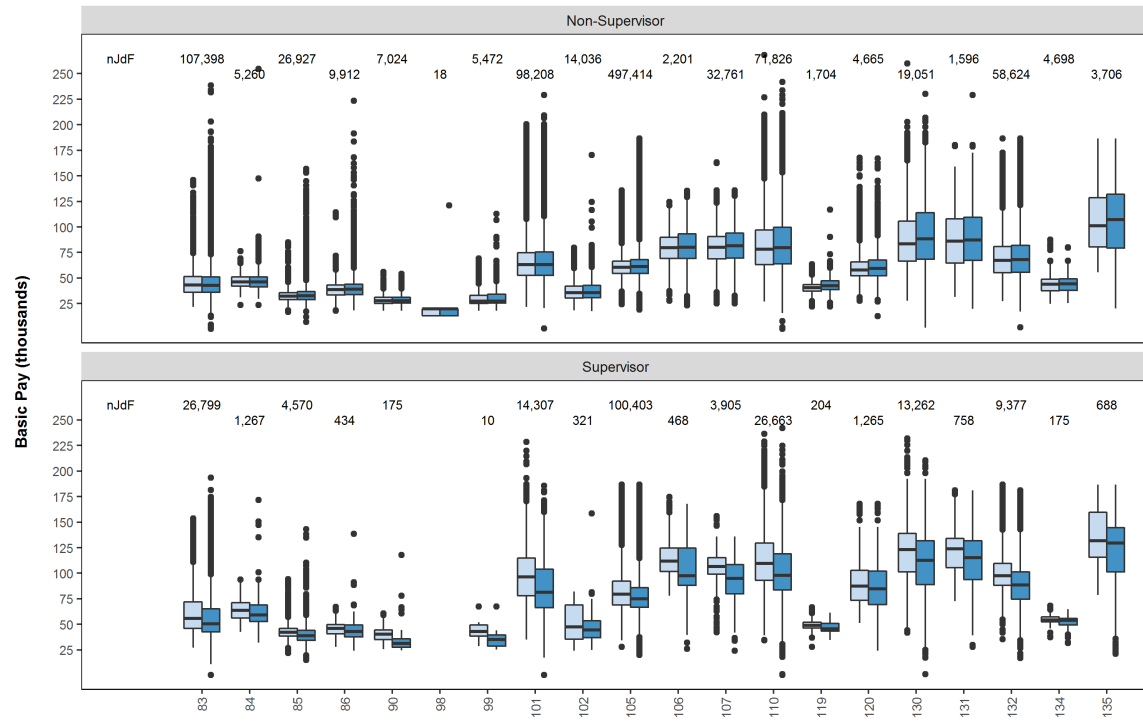
(b) Occupations 0083 through 0135

Figure 15: Basic pay distribution by occupation and supervisor status. All agencies combined. Authentic boxes on left, synthetic on right.
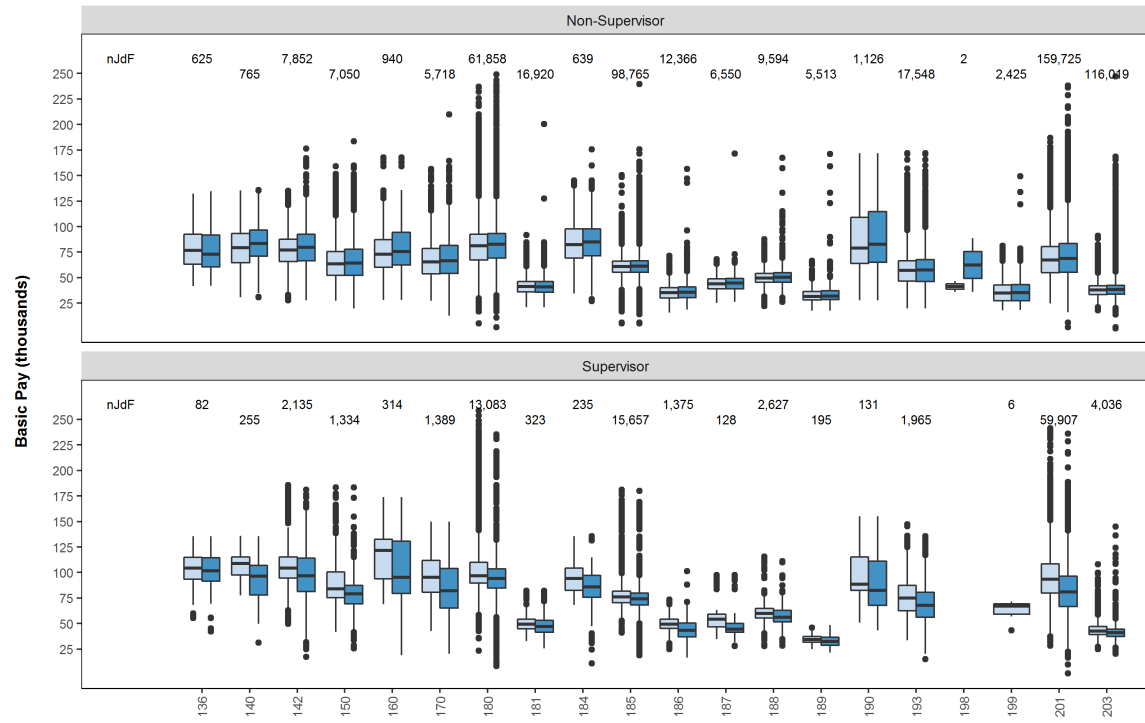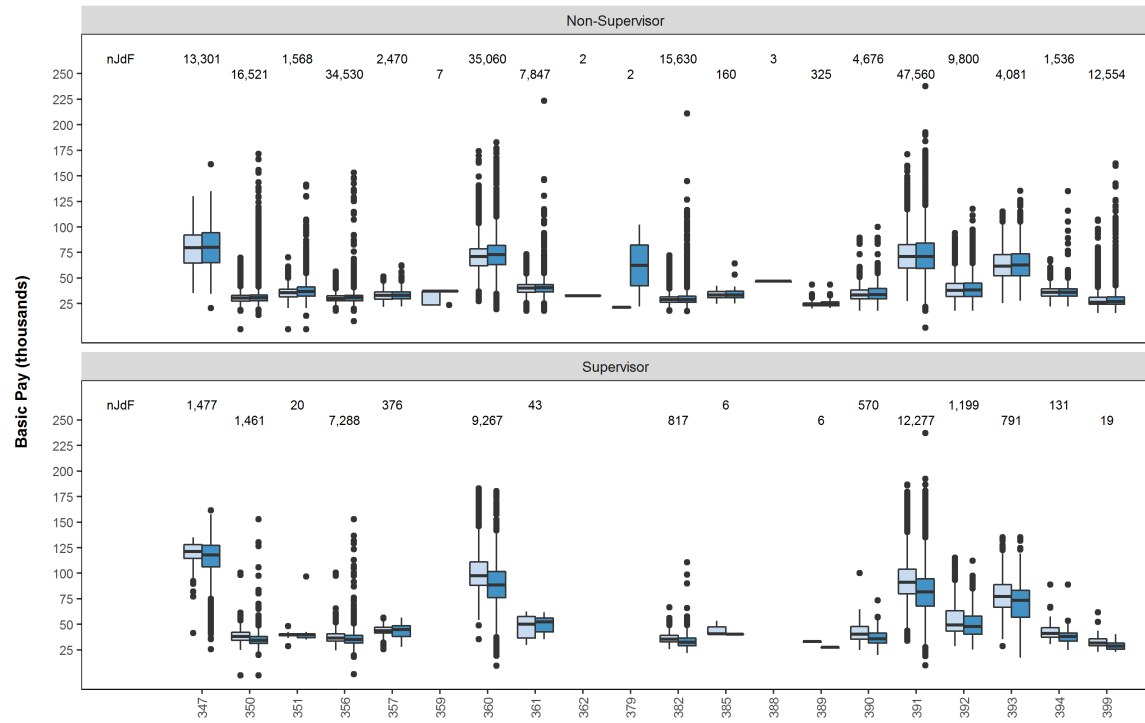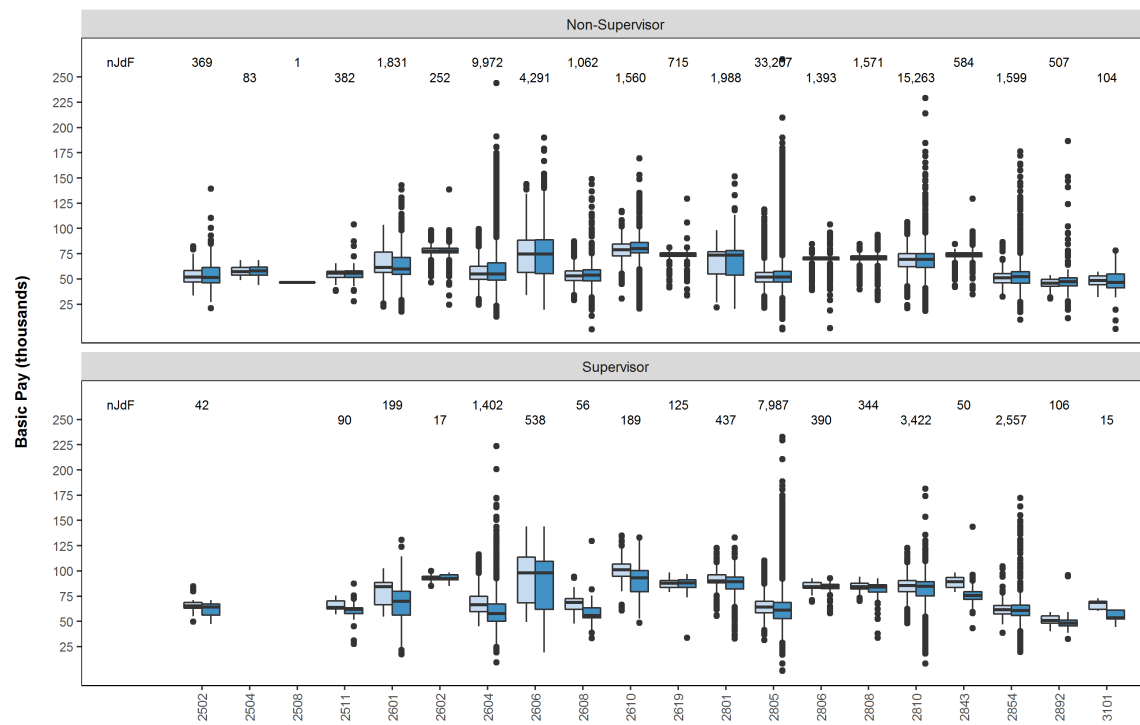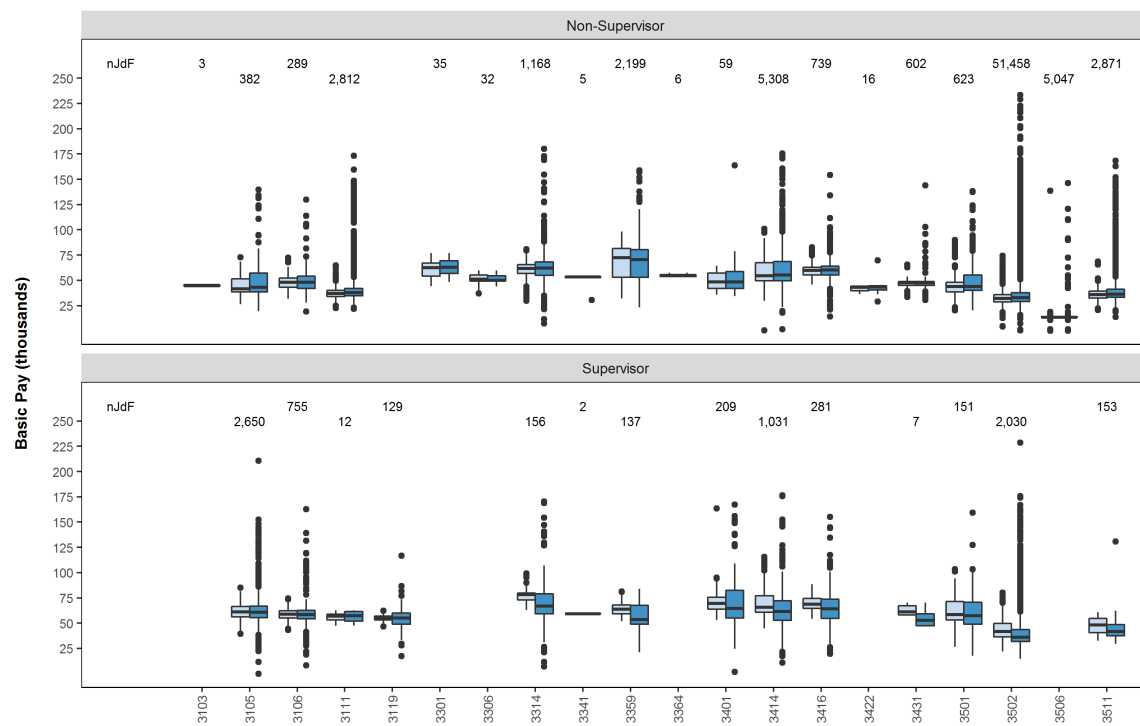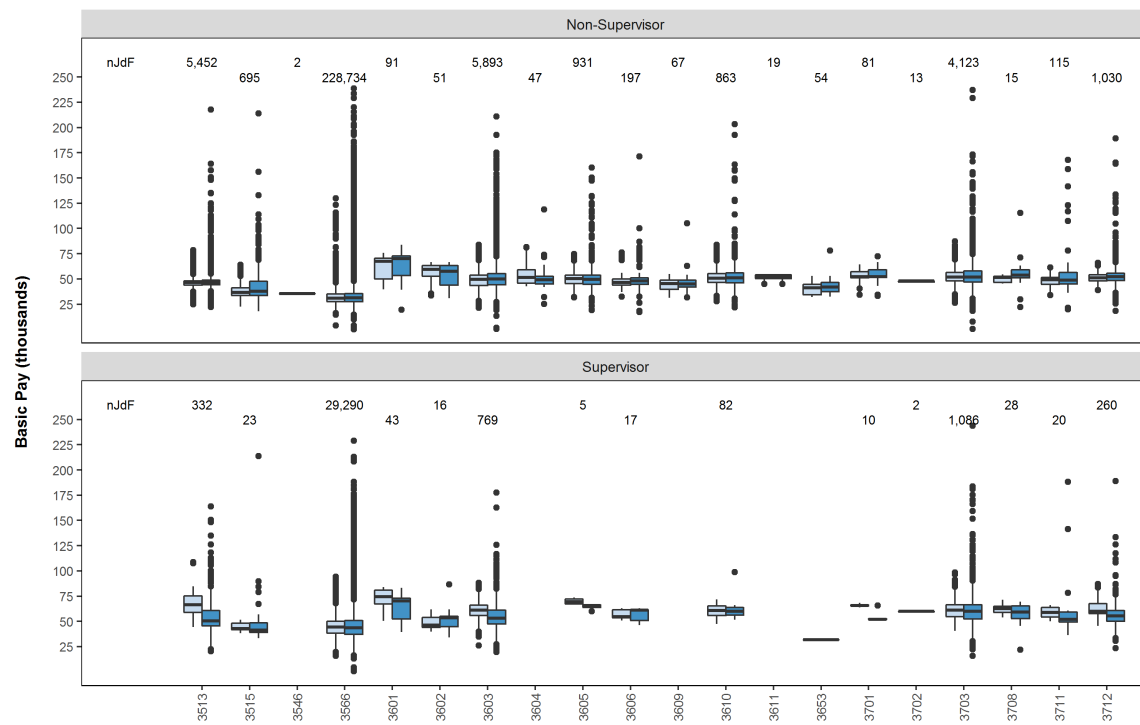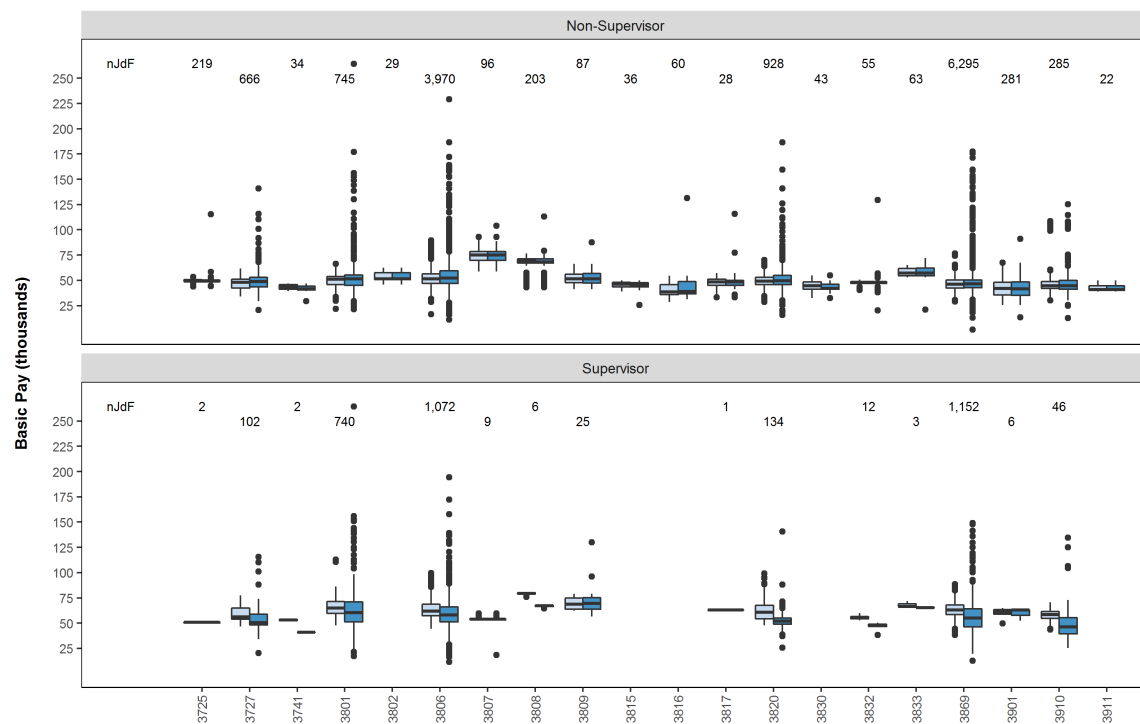
(a) Occupations 0136 through 0203
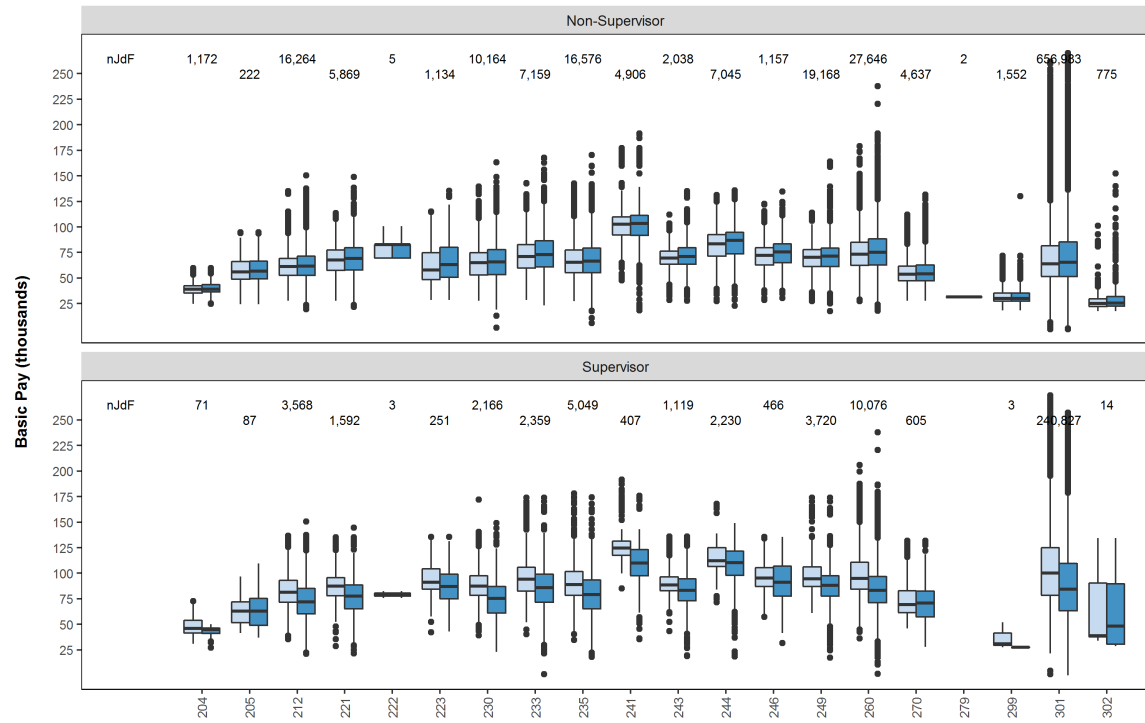
(b) Occupations 0204 through 0302

Figure 16: Basic pay distribution by occupation and supervisor status. All agencies combined. Authentic boxes on left, synthetic on right.
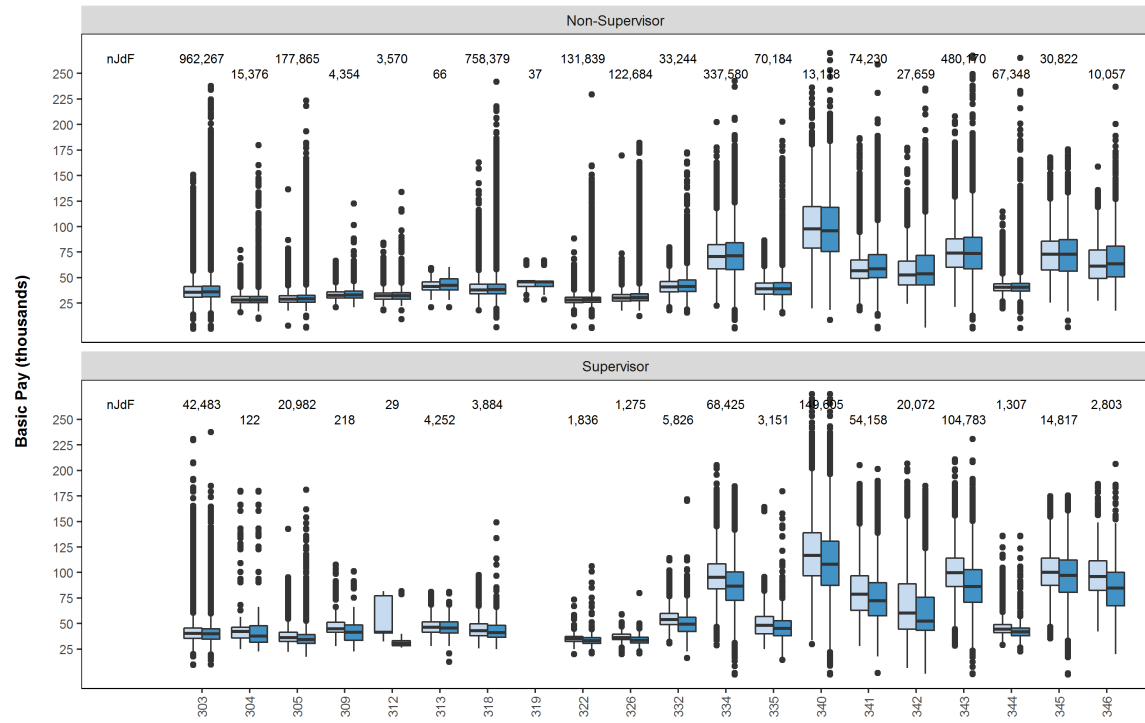
(a) Occupations 0303 through 0346

(b) Occupations 0347 through 0399

Figure 17: Basic pay distribution by occupation and supervisor status. All agencies combined. Authentic boxes on left, synthetic on right.

(a) Occupations 2502 through 3101 (trades)

(b) Occupations 3103 through 3511 (trades)

Figure 18: Basic pay distribution by occupation and supervisor status. All agencies combined. Authentic boxes on left, synthetic on right.

(a) Occupations 3513 through 3712 (trades)

(b) Occupations 3725 through 3911 (trades

Figure 19: Basic pay distribution by occupation and supervisor status. All agencies combined. Authentic boxes on left, synthetic on right.

MEAN LOG(BASIC PAY) BY GENDER, RACE, AND YEAR

The relationship of mean basic basic pay to joint combinations of sex, race, and year is important in human capital research and must be maintained in the synthetic data. Figure 20 plots mean log(pay) by year for females (on left) and males (on right) for races Native American (A), Asian (B), black (C), Hispanic (D), and white (E). Dashed lines for synthetic data, solid lines for authentic.

Observation: Differentiating colors may not be visible, but apparent pairings of lines (dashed near solid following similar trends) form race pairs. Although some systematic difference appears between data sets, inter-year and overall trends are very similar.
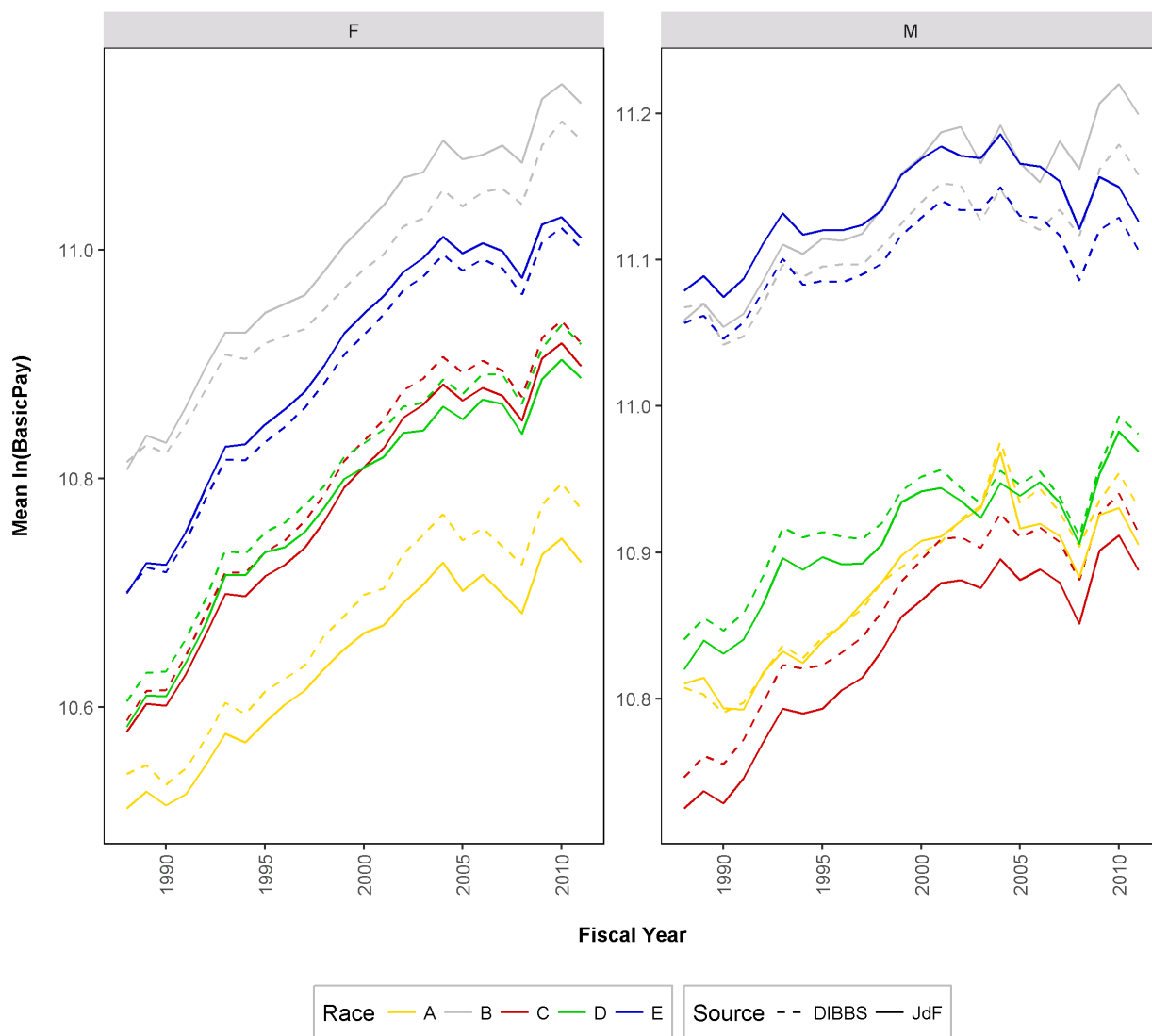
Figure 20: Mean basic pay by sex, race, and year. Female on left, male on right. Race codes: A = Native American, B = Asian, C = black, D = Hispanic, E = white. Dashed line for synthetic data, solid line for authentic.

GENDER PAY DIFFERENTIAL FIXED EFFECTS QUANTILE REGRESSION MODEL

Figure 21 plots the 0.1, 0.5, and 0.9 quantile estimates of difference in log(basic pay) between federal employee women and men by year, controlled for race, age, education, agency, and occupation.

Observation: Although some systematic separation appears between data sets, trends indicate similar proportion change with respect to time.
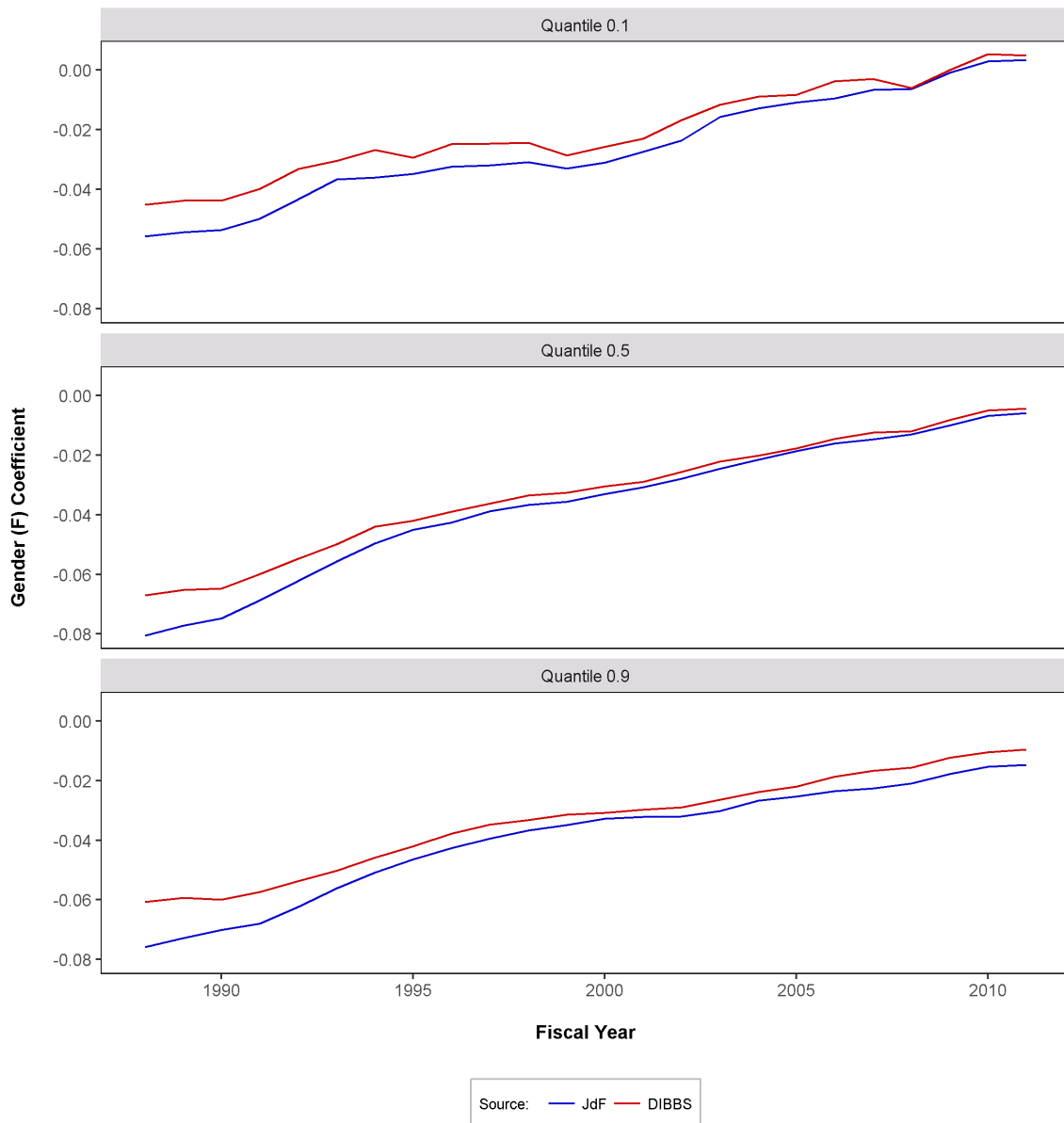

Figure 21: Quantile estimates from gender pay disparity fixed effects model. Upper line from synthetic data, lower line from authentic.

GENDER PROPORTION BY RACE, EDUCATION, AND YEAR

Proportion observations by gender is critical in models involving gender effects. Figure 22 plots proportion female employees by race, education, and year. Fitted lines are logistic regression models.

Observation: This four-way comparison (sex, race, education, and year) confirms good representation in synthetic data of gender proportion among important variable combinations in the authentic data. Fitted logistic regression models have nearly identical trends through fiscal years. Note the slight degradation in fit as observation count (n) decreases.
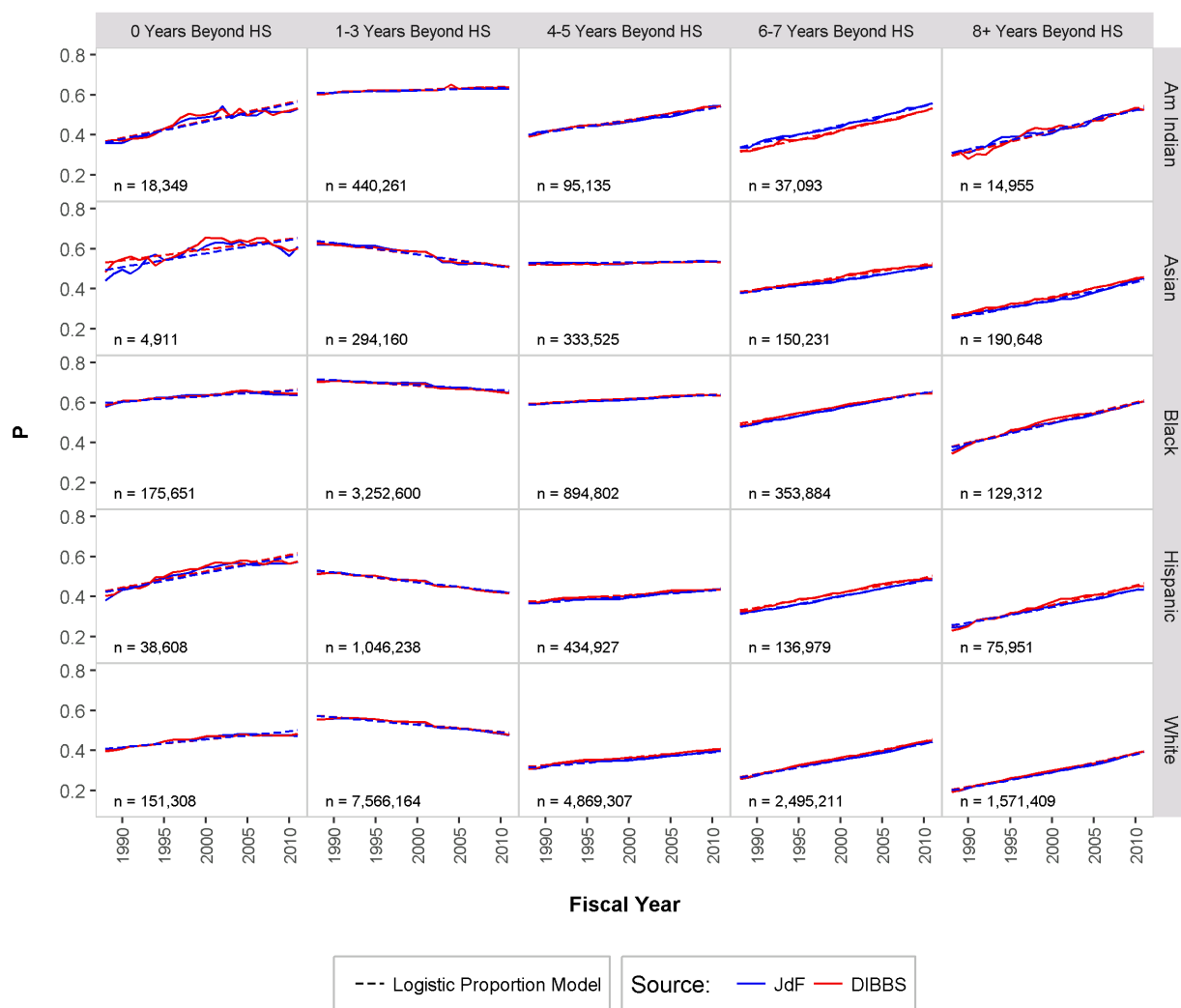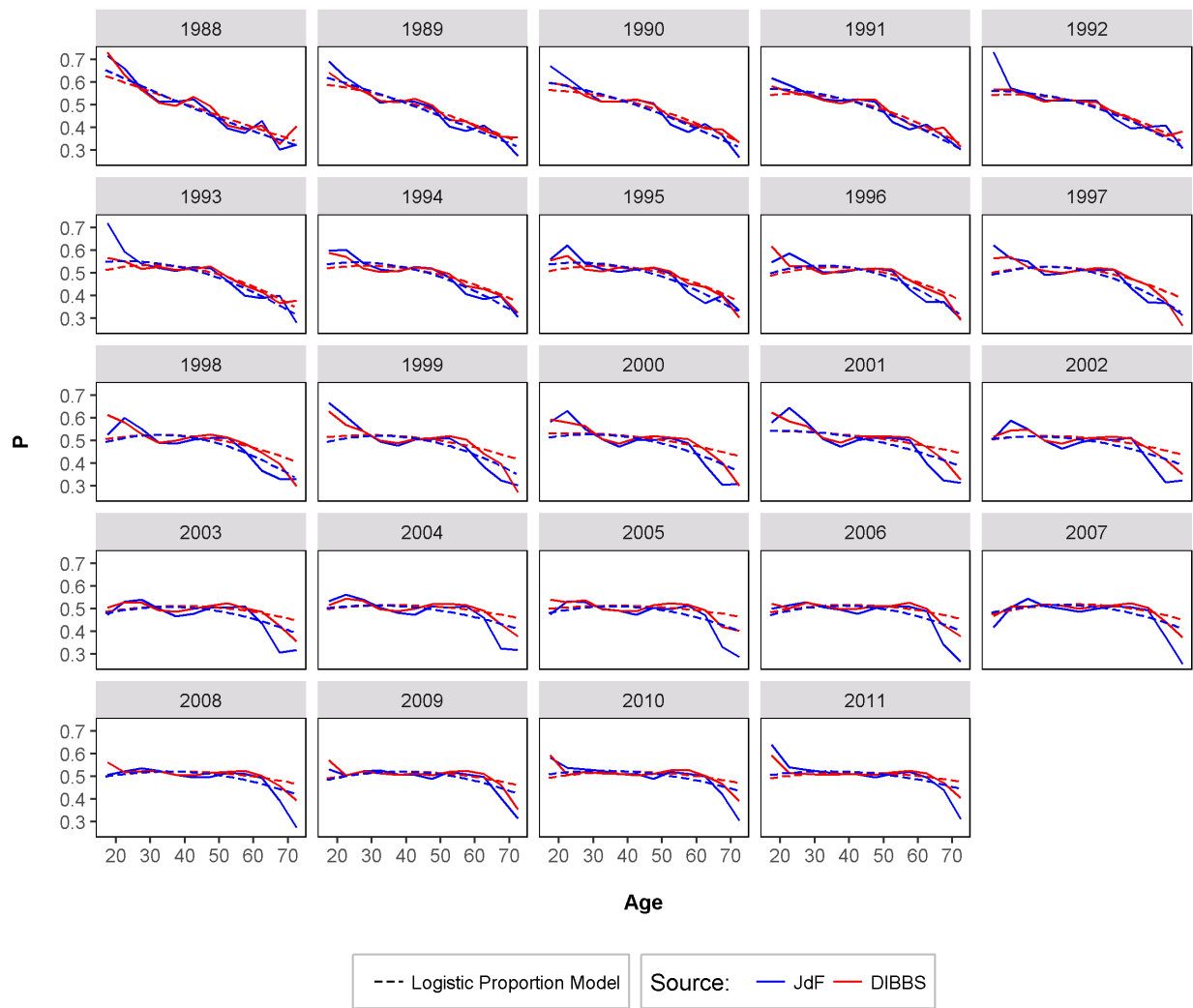

Figure 22: Proportion female observations by race, education, and year. Fitted lines are logistic regression estimates.

GENDER PROPORTION BY RACE, AGE, AND YEAR

Figures 23 through 27, show for each race, proportion female employees by age and year. Fitted lines are logistic regression models.

Observation: These four-way comparisons (sex, race, age, and year) confirm good representation in synthetic data of gender proportion among important variable combinations in the authentic data. Fitted logistic regression models have nearly identical trends through fiscal years.

Logistic models reveal agreement in trends between data sets.
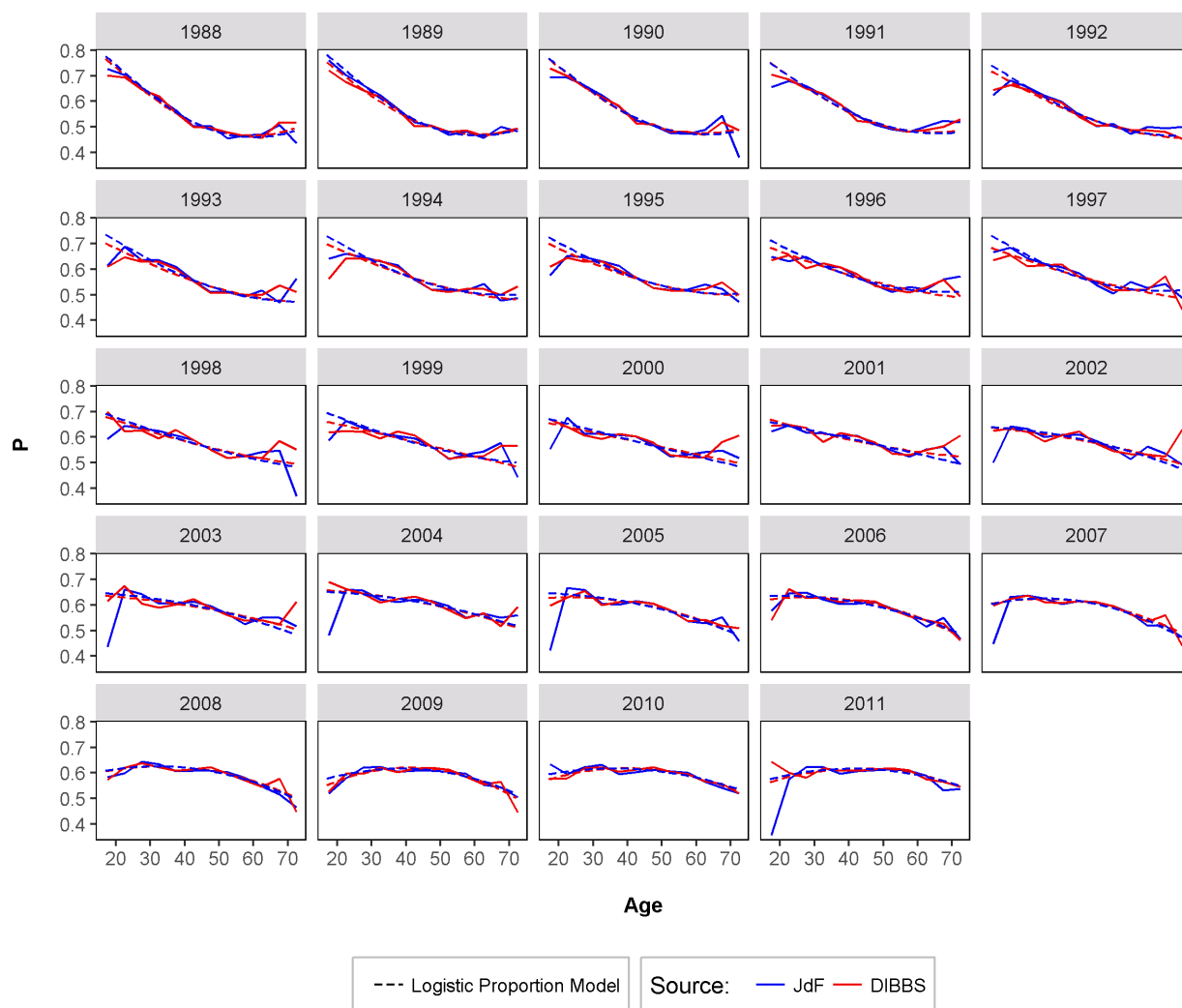
Figure 23: Proportion female observations by education and year. Race Native American. Fitted lines are logistic regression estimates.

Figure 24: Proportion female observations by education and year. Race Asian. Fitted lines are logistic regression estimates.
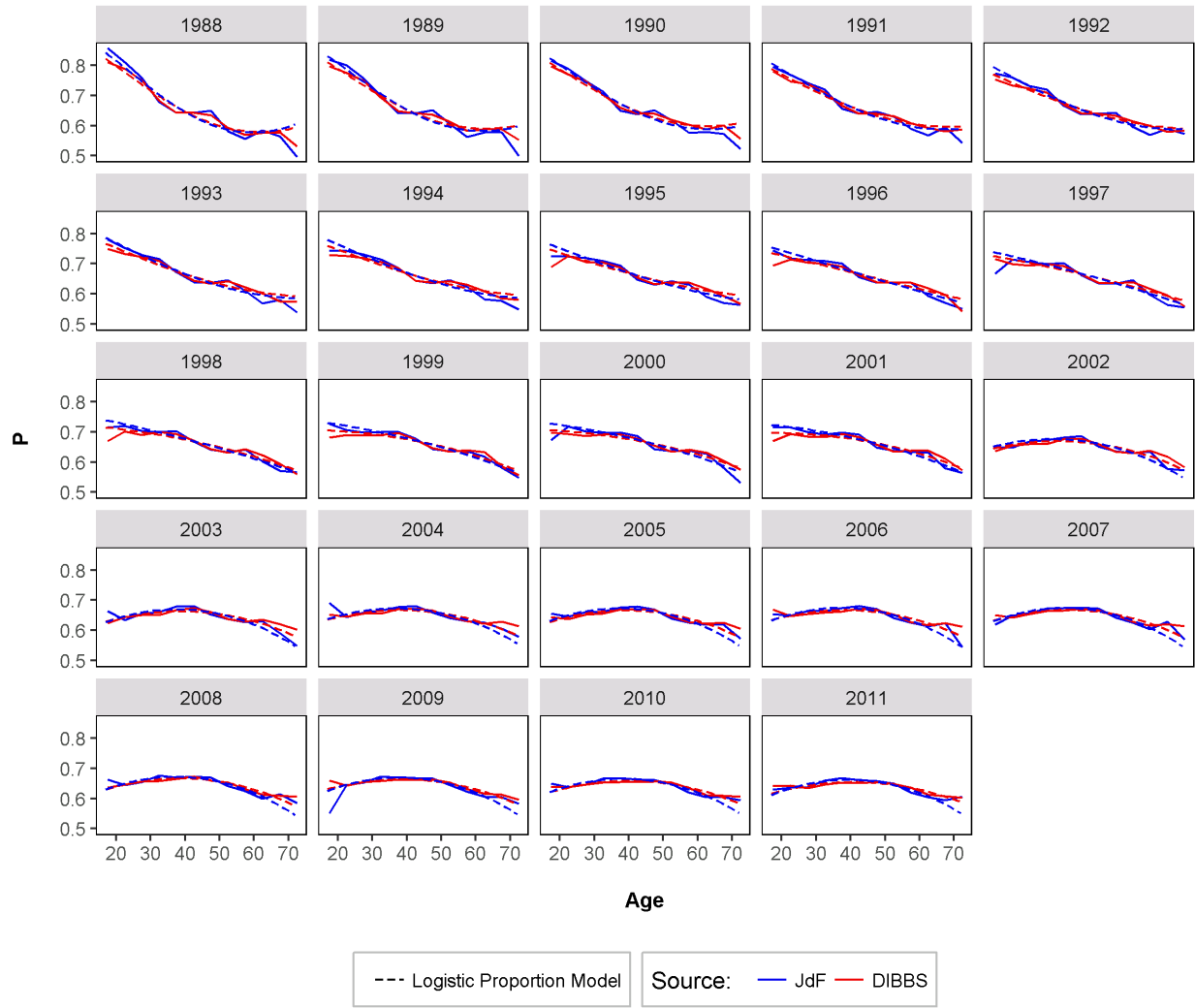
Figure 25: Proportion female observations by education and year. Race black. Fitted lines are logistic regression estimates.
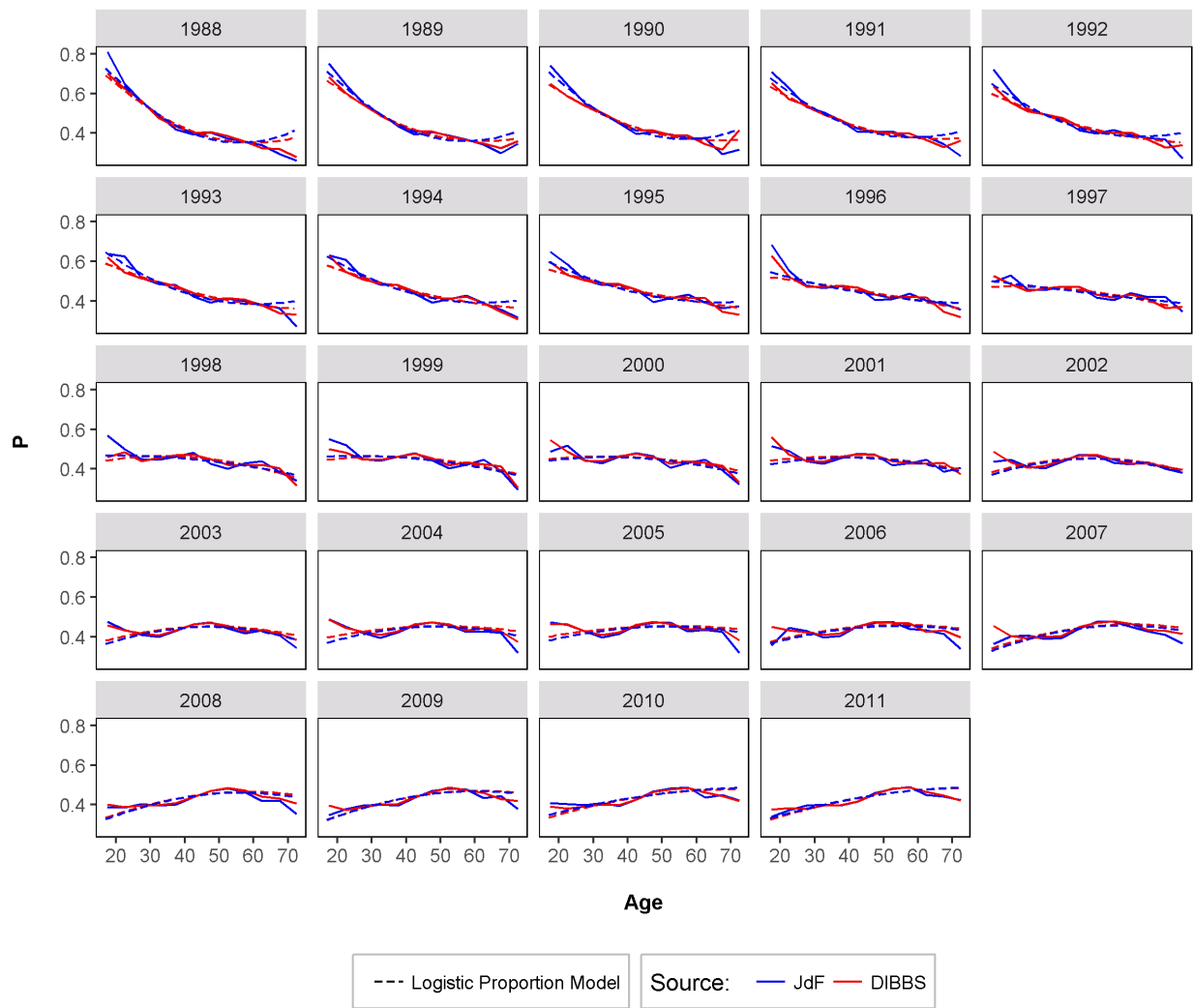
Figure 26: Proportion female observations by education and year. Race Hispanic. Fitted lines are logistic regression estimates.
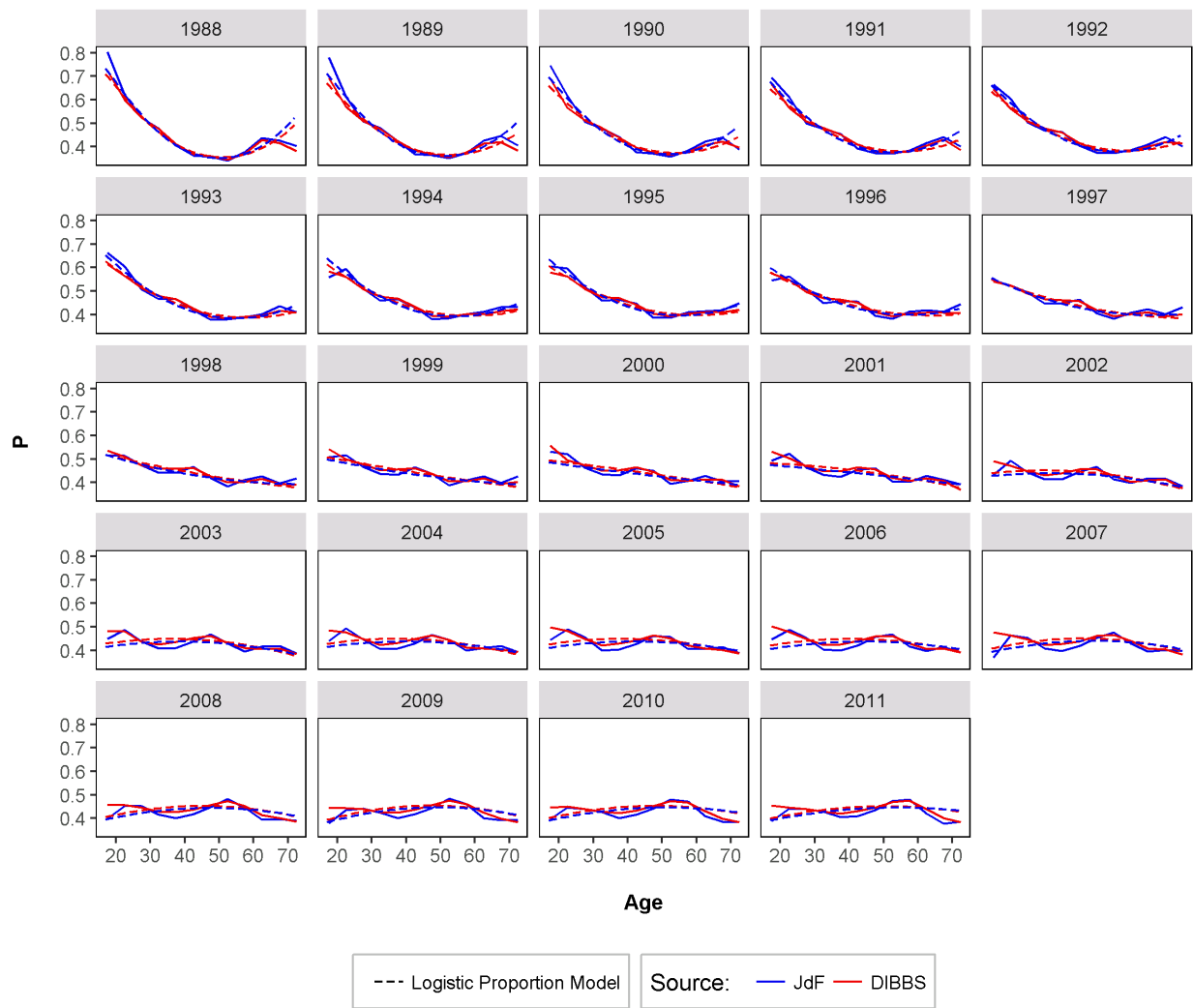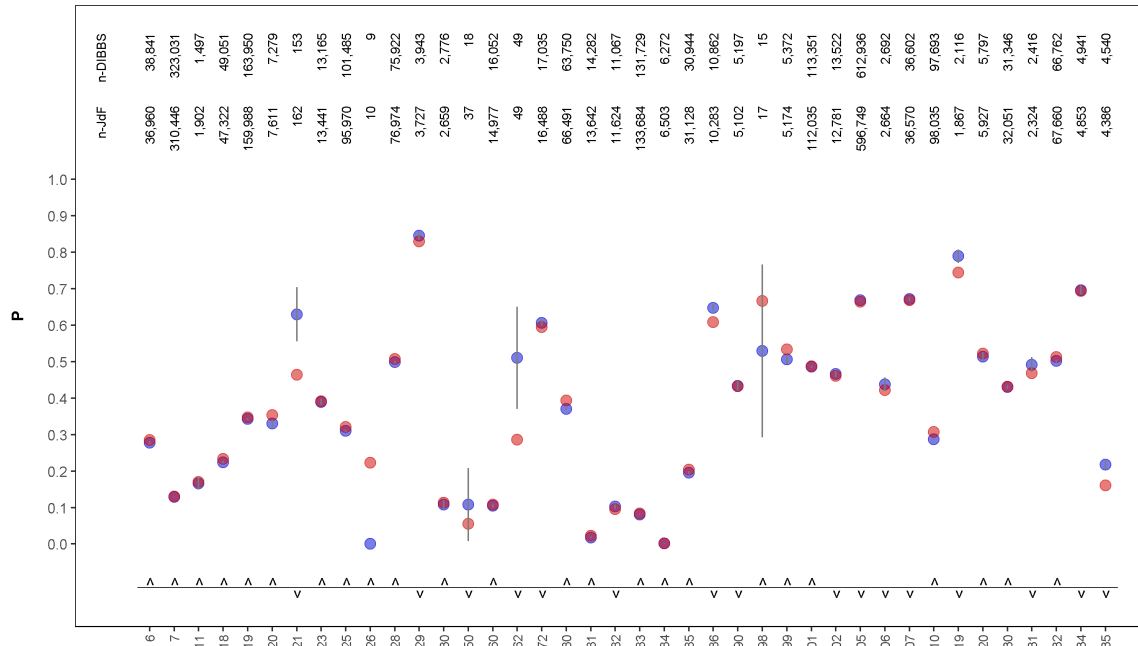
Figure 27: Proportion female observations by education and year. Race white. Fitted lines are logistic regression estimates.
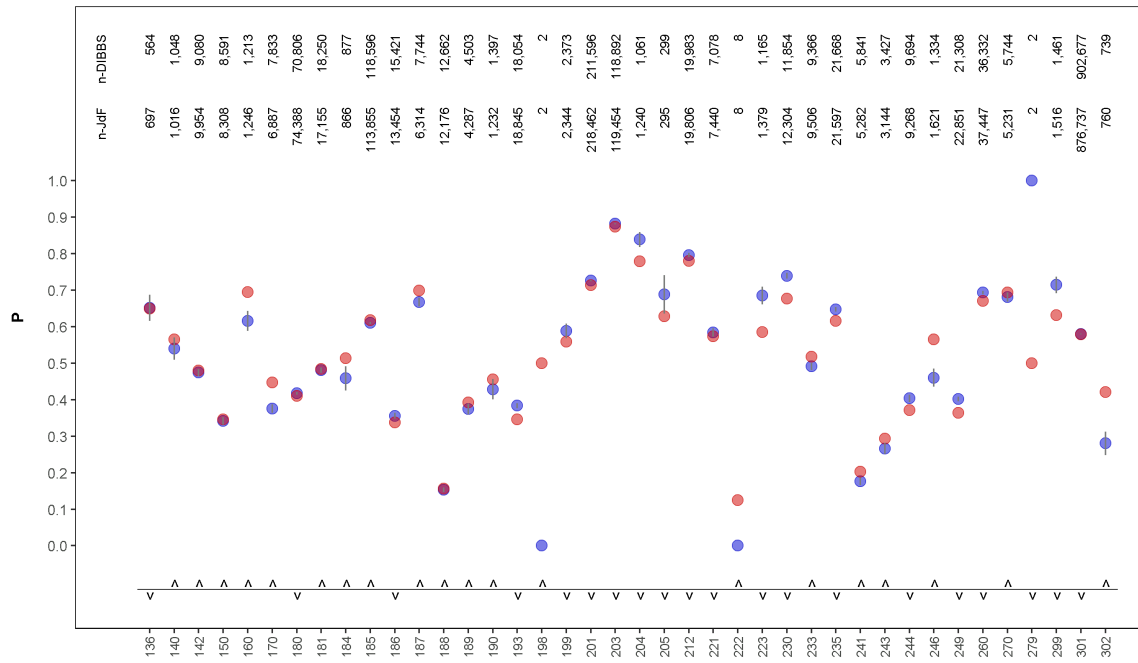
GENDER PROPORTION BY OCCUPATION

In the data supplied by OPM, 205 occupations (more than 25%) have proportion female observations below 0.05. 12 have proportion female observations greater than 0.95. Of these occupations, 148 have fewer than 1,000 total observations and 30 haver fewer than 10 observations. This presents challenges for accurate representation of authentic proportions in synthetic observations, while reducing the risk of individual employee identification. Figures 28 and 29 compare proportion female for the first 120 occupation codes. Figures 30 and 31 compare proportions for the first 120 trade occupations, which begin at code 2500. "n-DIBBS" indicates synthetic observation count, "n-JdF" indicates corresponding authentic observation count.

Observation: Agreement for large count occupations is indicated by proximity of points. Departure increases with decrease in observation count. Trade occupations, having generally low observation count and low proportion female in the authentic data, exhibit greater discrepancies than those observed in non-trade occupations.

(a) Occupations 0006 through 0135

(b) Occupations 0136 through 0302

Figure 28: Proportion female observations by occupation. All agencies combined. One synthetic and one authentic point per occupation.
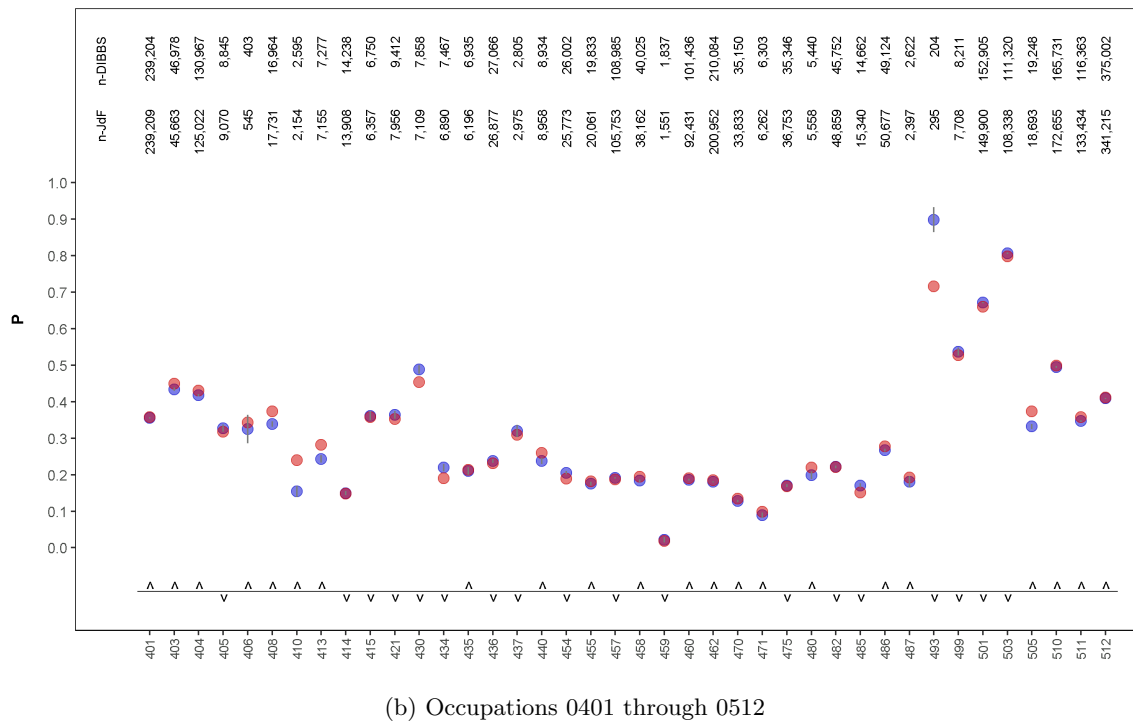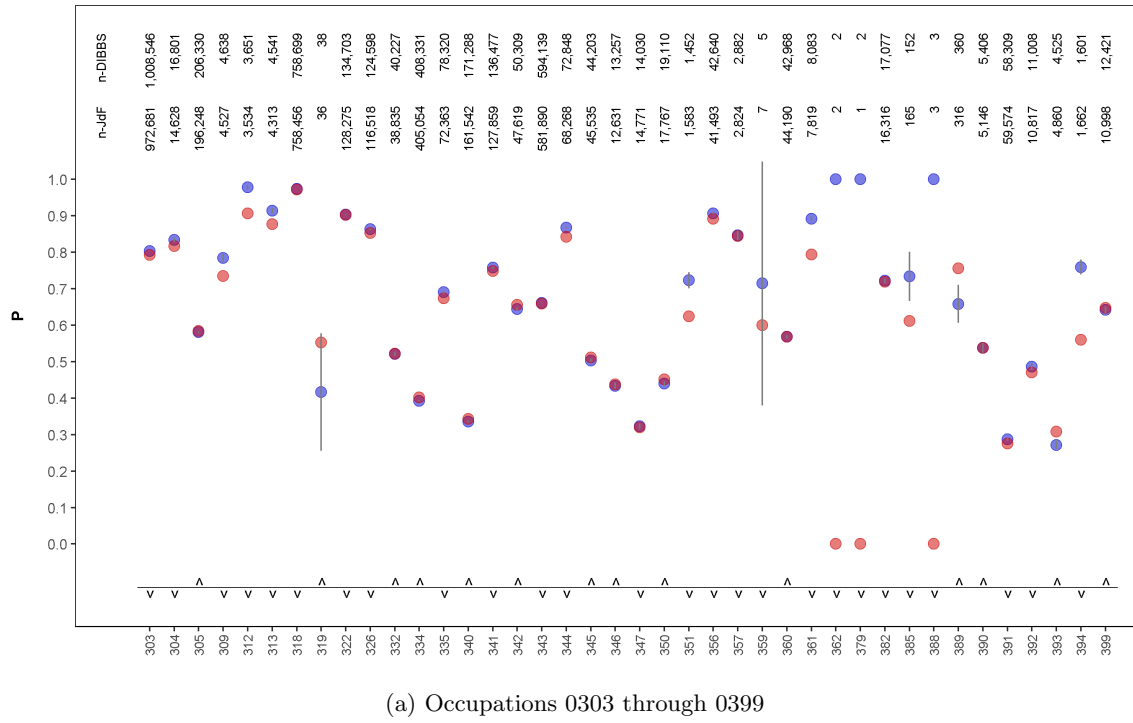
(a) Occupations 0303 through 0399
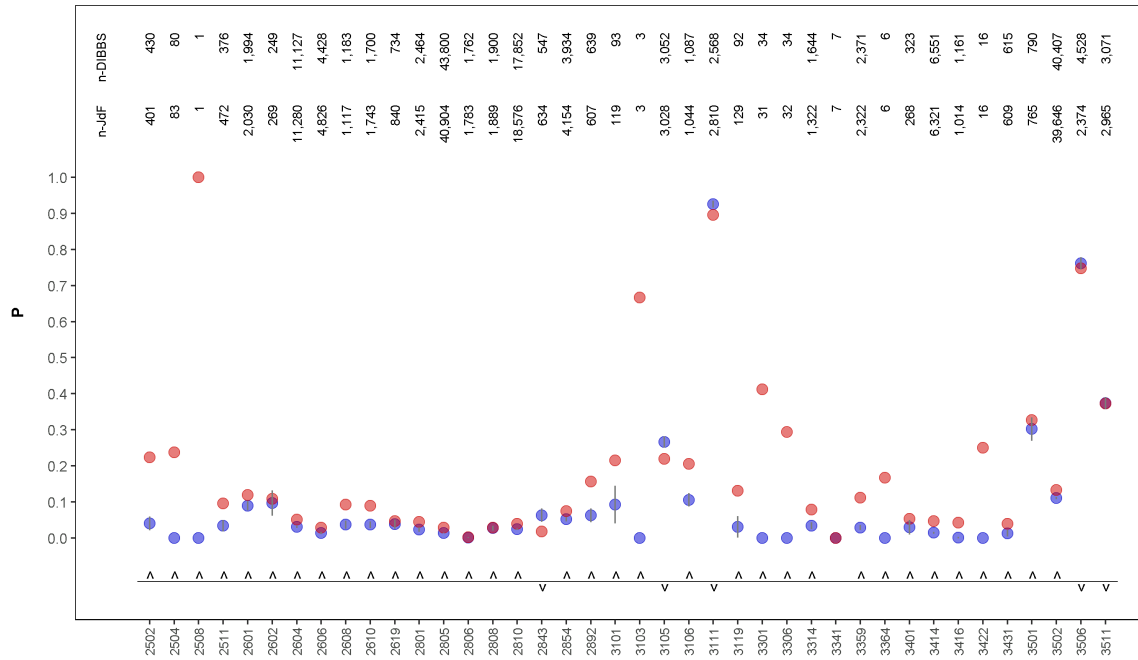

(b) Occupations 0401 through 0512

Figure 29: Proportion female observations by occupation. All agencies combined. One synthetic and one authentic point per occupation.
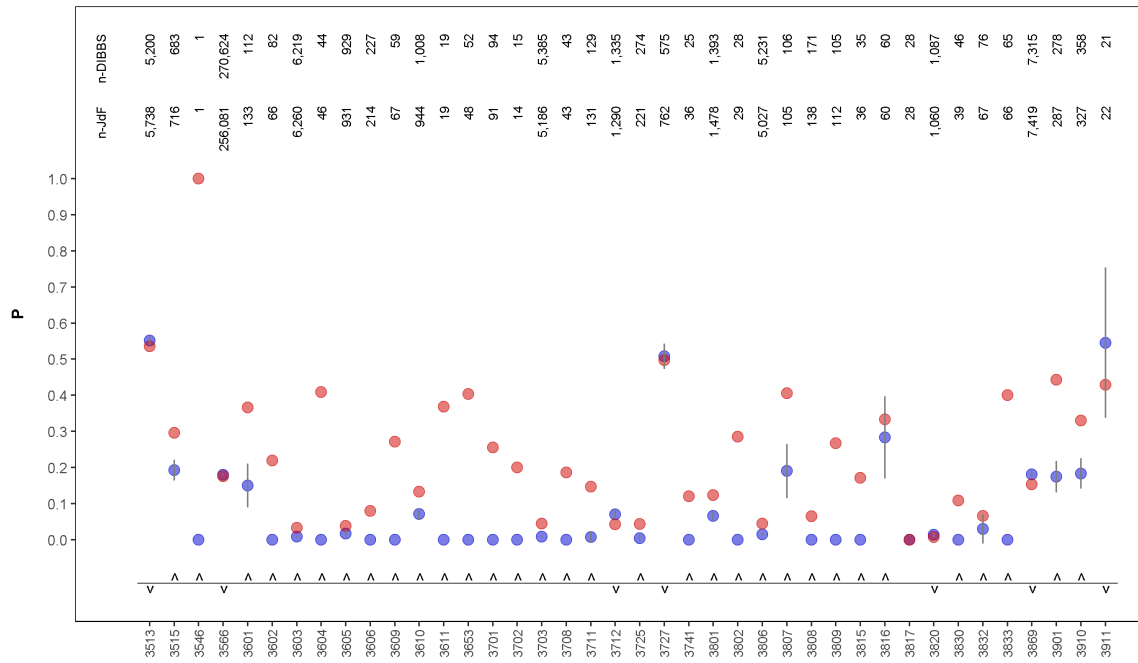
(a) Occupations 2502 through 3511 (trades)

(b) Occupations 3513 through 3911 (trades)

Figure 30: Proportion female observations by occupation. All agencies combined. One synthetic and one authentic point per occupation.

(a) Occupations 3940 through 4605 (trades)

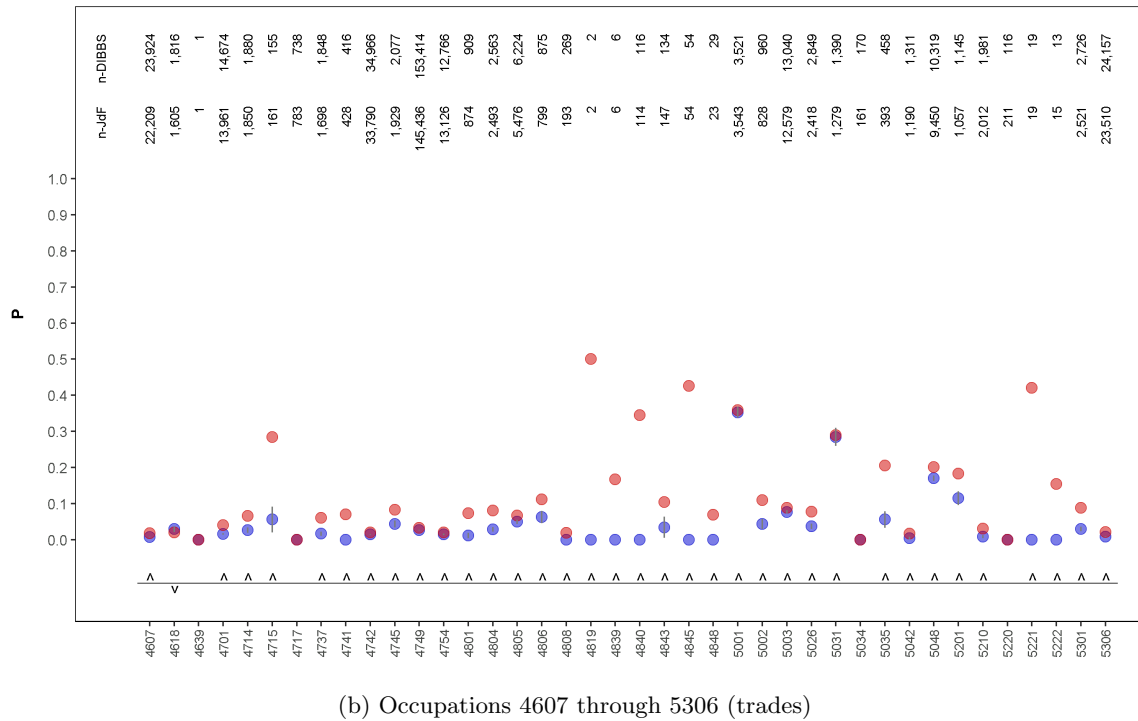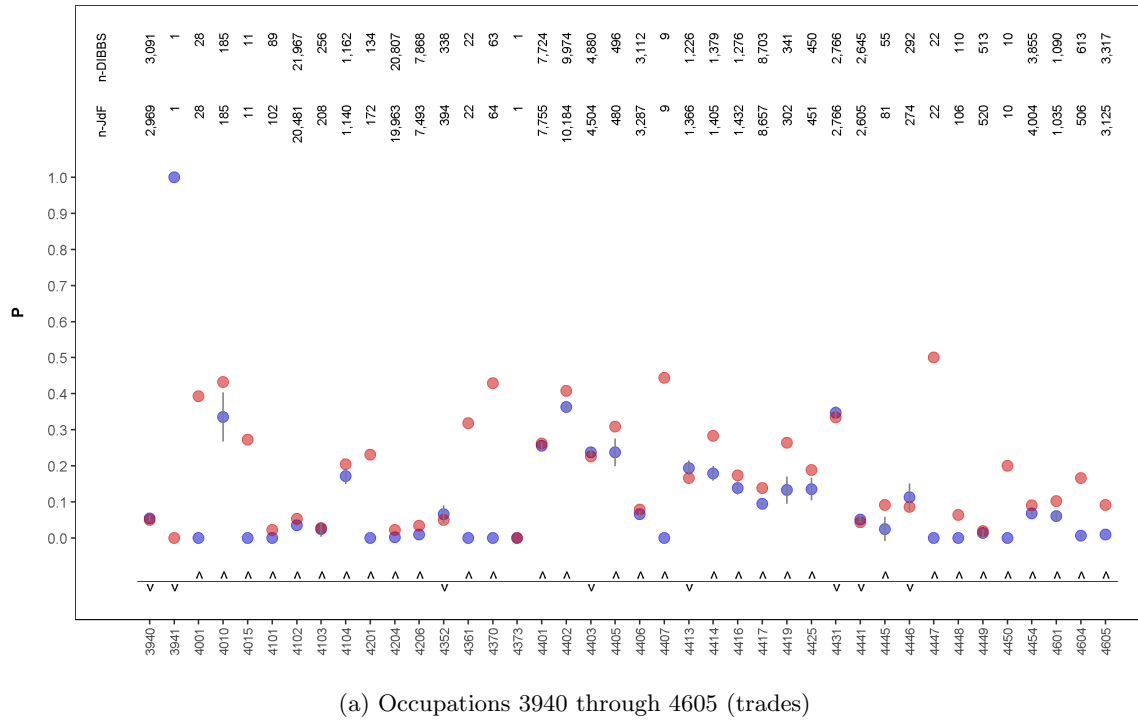(b) Occupations 4607 through 5306 (trades)

Figure 31: Proportion female observations by occupation. All agencies combined. One synthetic and one authentic point per occupation.

OCCUPATION GENDER PROPORTION KERNEL DISTRIBUTION

Figure 32 superimposes synthetic and authentic kernel density plots of proportion female employees by occupation.

Observations: Trade occupations in authentic data are slightly over-represented for proportions near zero and above 0.75. There is slight discrepancy in synthetic proportions near 0 and above 0.75, which is compensated for near central proportions.
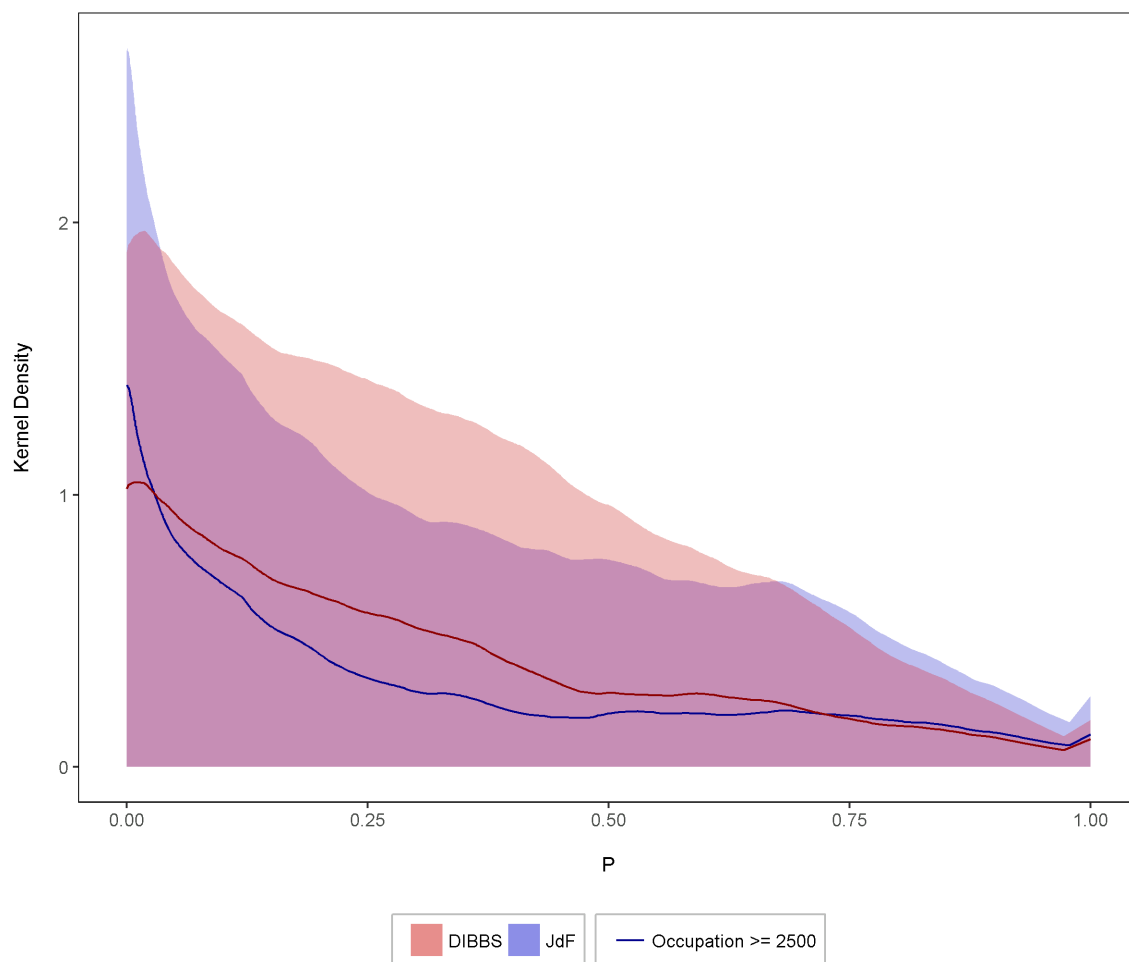

Figure 32: Occupation proportion female kernel density. Synthetic and authentic distributions superimposed. Trade occupations slightly over-represented near zero and above 0.75 in authentic data. Slight discrepancy in synthetic proportions at extremes. Compensated for near central proportions.

GENDER PROPORTION LOGISTIC REGRESSION CLASSIFIER FOR TRADE OCCUPATIONS

A gender classifier for occupations with code greater 2500 (trades) using the logistic regression model

$$\hat{\boldsymbol{p}} = f(\hat{\boldsymbol{\beta}}_{race}race + \hat{\boldsymbol{\beta}}_{age}age + \hat{\boldsymbol{\beta}}_{age^2}age^2 + \hat{\boldsymbol{\beta}}_{ed}ed + \hat{\boldsymbol{\beta}}_{ed^2}ed^2 + \hat{\boldsymbol{\beta}}_{occ}occ),$$

where $f()$ estimates proportion female observations by race, age, education, and occupation, was used to classify sex, such that all observations associated with combinations of independent variables with $\hat{\boldsymbol{p}} \geq 0.5$ are classified as female. Figure 33 plots, for fiscal years 1988-2011 (all years supplied by OPM) and $\hat{\boldsymbol{p}}$ values from 0 to 1.0, the proportion of accurate female observation classification (y-axis) against the proportion accurate male classification (x-axis).

Observation: Although the classifiers, exhibiting somewhat flat ROC curves (near the $\hat{\boldsymbol{p}}$=0.5 reference line of slope 1.0) appear to be of limited utility, those derived from synthetic data are nearly identical their counterparts derived from authentic data, including an apparent reduction in utility (nearer to the reference line) as fiscal years advance. Incidentally, the reduced utility with year may reflect structural changes in proportion female for trade occupations during this period.
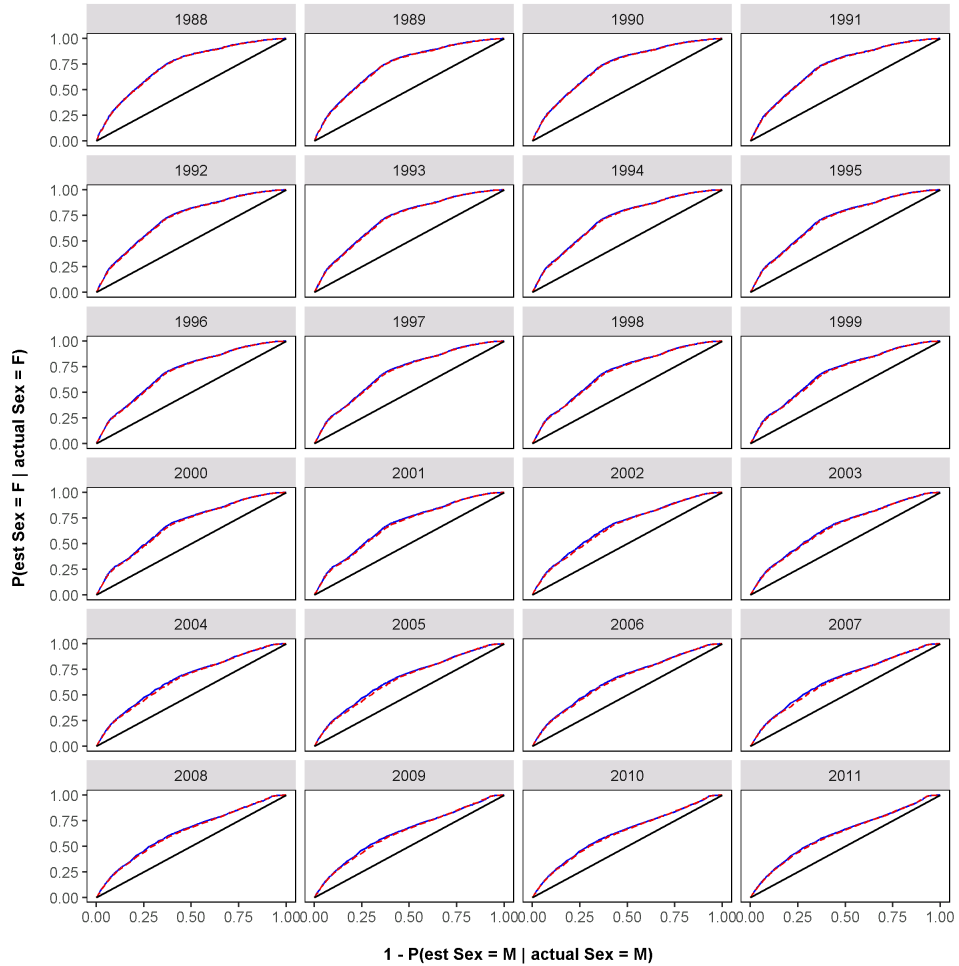

Figure 33: Proportion female race, age, education, occupation classifier ROC curves. One curve per data set per by fiscal. Agreement in classifier accuracy indicated by overlapping curves. Pattern of decreased accuracy as years progress captured in both data sets.

The next sections (each beginning with "THE RISE OF GRADE IN THE FED GOVT") use results from Bolton and de Figueiredo (2016). Each compares the fit of actual research models to corresponding sub-sets of synthetic and authentic data. Some figures include graphs that were constructed using corresponding data from OPM's on-line FedScope data repository (U.S. Office of Personnel Management).

THE RISE OF GRADE IN THE FED GOVT: WAGE BILL DECOMPOSITION

Figure 34 shows the annual change in the total federal employee wage bill categorized by source: change in grade, change in step rate, and other changes. Synthetic data are represented by a dashed line, authentic data by a solid line.

Observation: Although highly aggregated, each graph is informative and the synthetic data provide near identical insight into overall wage change patterns as do the authentic data.
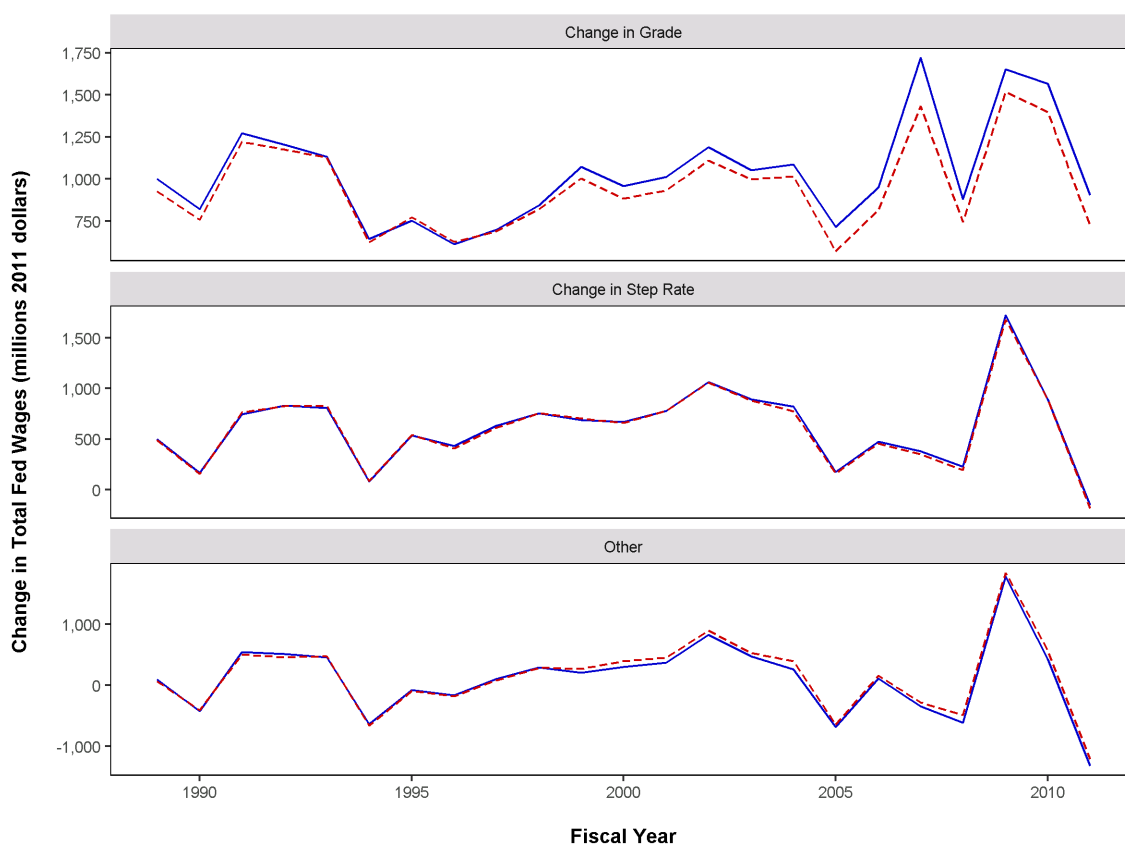

Figure 34: Change in U.S. federal government total wage bill. Millions of 2011 dollars by change in grade, change in step rate, and other changes. Fiscal years 1988 through 2011.

The GS pay plan represents approximately 80% of the observations in the data provided by OPM. Accordingly, change in distribution of grade within this pay plan is an important consideration when conducting human capital research with these data. Figure 35 shows the change in grade distribution from fiscal years 1988 (solid line) to 2011 (dashed line).

Observation: Although highly aggregated, each graph is informative and the synthetic data provide near identical insight into overall change patterns as do the authentic data.
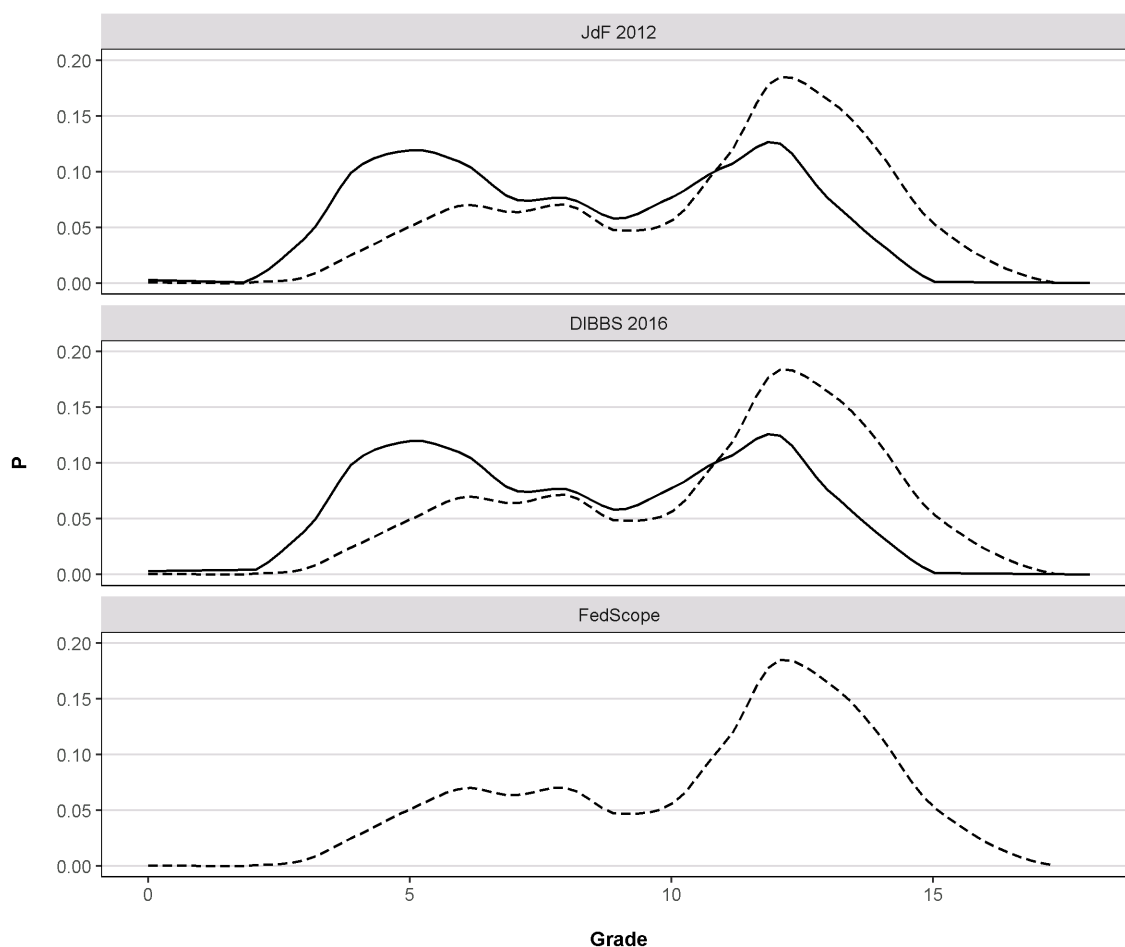

Figure 35: Change in GS grade distribution. Fiscal years 1988 (solid line) and 2011 (dashed line). Near identical distribution in synthetic and authentic data.

THE RISE OF GRADE IN THE FED GOVT: 90/10 PAY PERCENTILE RATIO

In addition to a general increase in wages over their study period, Bolton and de Figueiredo show, for the GS pay plan, that wages for employees at the top end increased at a greater rate than for lower paid employees. Figure 36 plots the ratio of 90th and 10th basic pay percentiles for the GS pay plan, grade less than or equal to 15. Synthetic data are represented by a dashed line, authentic data by a solid line.

Observations: The rise of nearly 0.2 measured in the authentic data is informative and, along with very close tracking of local trends throughout the period, is apparent in the synthetic data.
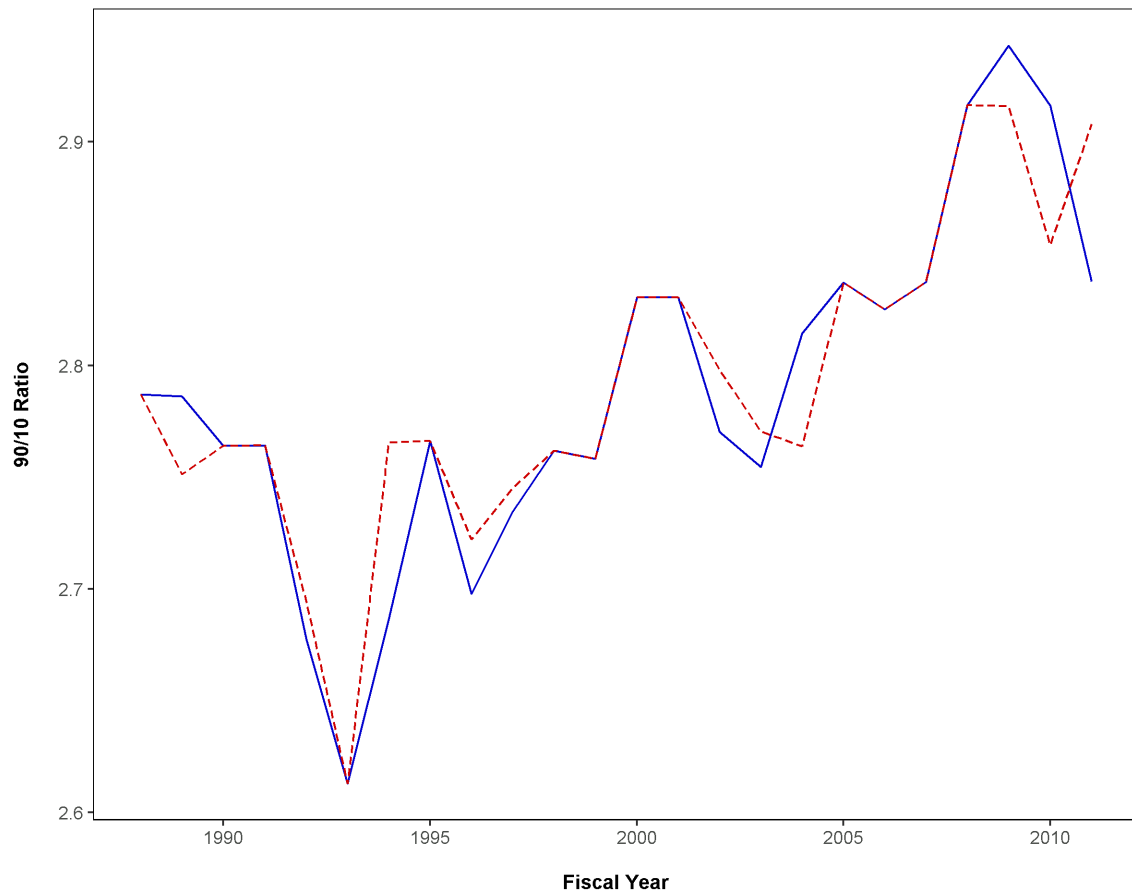

Figure 36: Ratio of 90th and 10th basic pay percentiles by year. GS pay plan, grade *leq* 15. Synthetic data dashed line, authentic data solid.

THE RISE OF GRADE IN THE FED GOVT: BASIC PAY QUANTILE REGRESSION

Ordinary least squares regression estimates expected values, given levels of independent, or predictor, variables. Of interest with distribution of income are quantiles, or an estimate pay below which a given proportion of observations are observed. Figure 37 plots, for pay plan GS, grade less than or equal to 15, the slopes estimated from the simple linear quantile regression of the logarithm of basic pay on fiscal year (1988-2011). These slopes represent change in corresponding quantile per year. One model fit for each quantile from 0.1 through 0.9 in 0.1 increments. Synthetic data are represented by a dashed line, authentic data by a solid line.

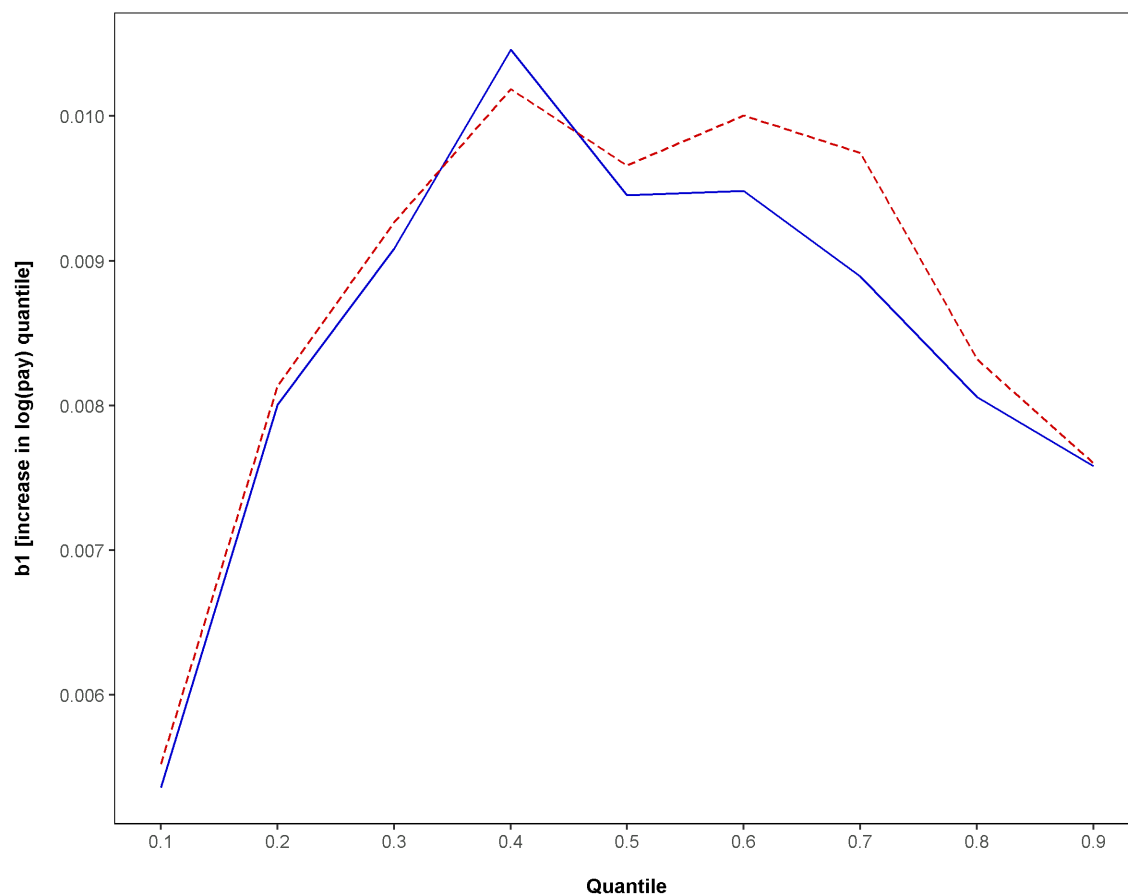Observation: Similar trends in slope of log(pay) quantile with respect to year are revealed by both data sets.

Figure 37: Coefficients (change per year) from quantile regression of log(pay) on year. GS pay plan, grade ≤ 15. 1988-2011. Synthetic data dashed line, authentic data solid.

References

Alexander Bolton and John M. de Figueiredo. Why have federal wages risen so rapidly? Tech. rep., Duke University Law School, 2016.

Alexander Bolton and John M. de Figueiredo. Measuring and explaining the gender wage gap in the U.S. federal government. Tech. rep., Duke University Law School, 2017.

U.S. Office of Personnel Management. FedScope. URL https://www.fedscope.opm.gov/datadefn/aehri_sdm.asp.

U.S. Office of Personnel Management, Data Analysis. Data, Analysis, and Documentation. URL https://www.opm.gov/policy-data-oversight/data-analysis-documentation/.

U.S. Office of Personnel Management, GDS. Guide to Data Standards. URL https://catalog.data.gov/dataset/guide-to-data-standards-gds.