Providing Access to Confidential Research Data Through Synthesis and Verification: An Application to Data on Employees of the U.S. Federal Government

Supplement: Synthetic Data Validation

Duke University Synthetic Data Project

March 18, 2018

The following is excerpted from work done as part of the Synthetic Data Project 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 variables important to human capital research 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, A). For additional information and guidance on use and interpretation of data made available by OPM, see (U.S. Office of Personnel Management, B).

[1] A complete description of both data sets and sources is available in the main document that the current document supplements.

[2] "JdF" is nomenclature for a particular FOIA request that resulted in receipt of authentic data from OPM, which was used to generate synthetic data.

This document is organized in sections, each addressing a particular validation of univariate distribution, covariate distribution, or fit of research models to sub-sets of data.

List of Sections

1 Data Set Overview

Authentic data supplied by OPM consists of what are called status records and contain important human capital indicators, such as agency employed by, grade, sex, race, age, occupation, education, and pay as of September 30th of each year reported. The OPM data set includes status records for select agencies from years 1988 through 2011. The synthetic data were generated using covariate and longitudinal relationships observed in the authentic data. Table 1 gives a basic outline of the authentic and synthetic data sets.

Table 1: Outline of authentic and synthetic data sets

	Authentic		Synthetic	
Total observation count	28,257,629		28,512,573	
Fiscal years represented	1988-2011		1988-2011	
Distinct employees	3,511,824		3,510,406	
Distinct agencies	607		604	
Distinct occupations	803		803	
Records with basic pay = 0	49,696	1.4%	55,405	1.6%
Duplicate employee, year combinations	21,126	0.6%	0	0.0%
Year, agency combinations absent in other set	15	0.2%	0	0.0%
Year, agency, occupation combinations absent in other set	2156	0.6%	0	0.0%
Remaining variable codes are as they appear in the GDS				

Notes:

- Based on discussions with OPM, basic pay should never be zero. Records with such values are considered invalid and are excluded from certain of the analyses that follow, particularly those that categorize or use basic pay as a response variable. Records with basic pay equal to 0 were generated in the synthetic data in order to simulate actual properties of the authentic data.
- According to OPM, employee ID and year should be unique. However, in the authentic data, a small proportion of duplicates are observed. Certain of the following analyses restrict observations to a unique employee ID, year combinations. In selecting distinct combinations, preference is given to records with non-zero basic pay. If multiple records remain for an employee, year combination after basic pay filtering, a single record is randomly selected for the employee and year. Synthetic records do not have duplicated ID, year combinations.
- A small proportion of year, agency and year, agency, occupation combinations appearing in the authentic data do not appear in the synthetic data. However, all combinations appearing in the synthetic data are reflected in the authentic data. Thus, the synthetic data do not create new agency, occupation, time relationships.

2 Covariate Relationships

For the synthetic data to have utility, covariate relationships must reflect those observed in the authentic data. This section compares, for significant human capital variables, synthetic and authentic two variable correlation and correlation of primary variables with two variable interactions. Data are limited to pay plan GS, full-time observations, which represents the largest federal white collar pay plan and account for approximately 75% of observations supplied by OPM.

2.1 Two Variable Correlation

Figure 1 shows 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.

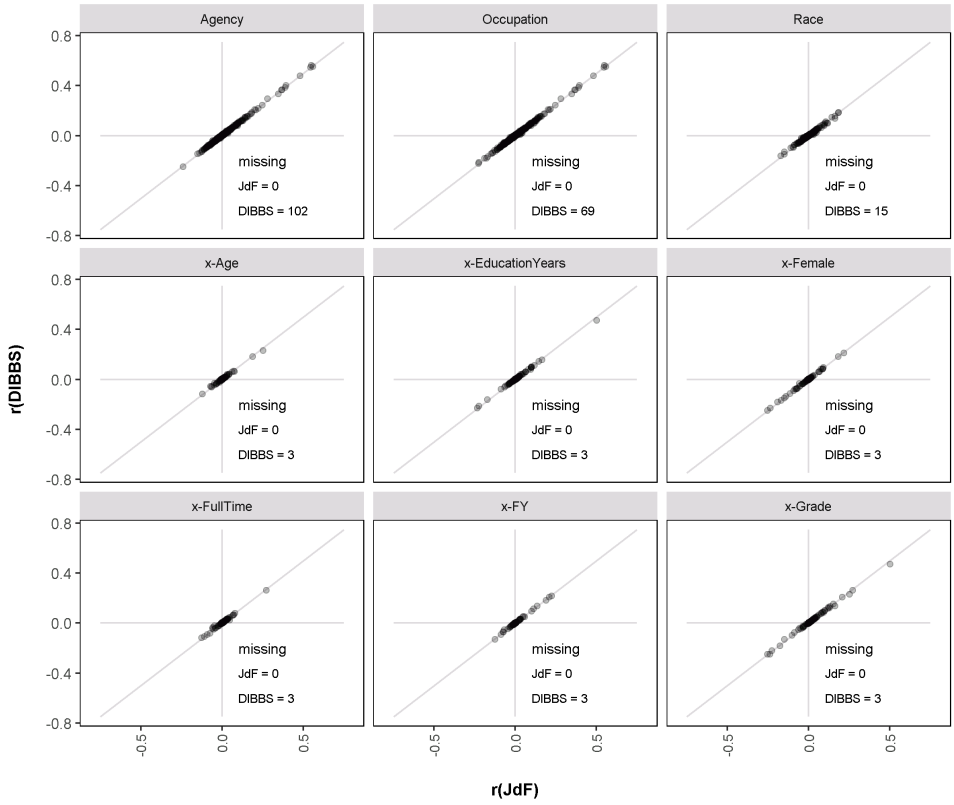


Figure 1: Two variable correlations of corresponding levels of synthetic and authentic data. Synthetic level correlation on y-axis, corresponding authentic correlation on x-axis.

2.2 Correlation of Primary Variables With Two-Variable Interactions

Figures 2 through 6 show 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 the slope 1.0 reference line indicates agreement of three-variable associations between data sets and implies depth of utility beyond simple pairwise relationships.

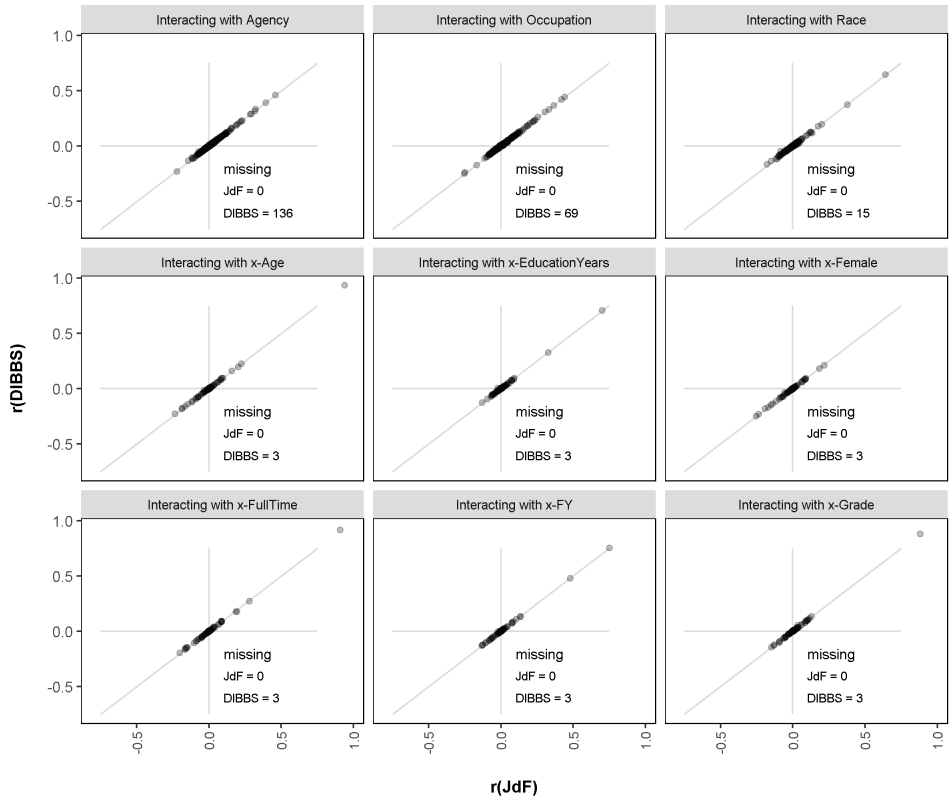
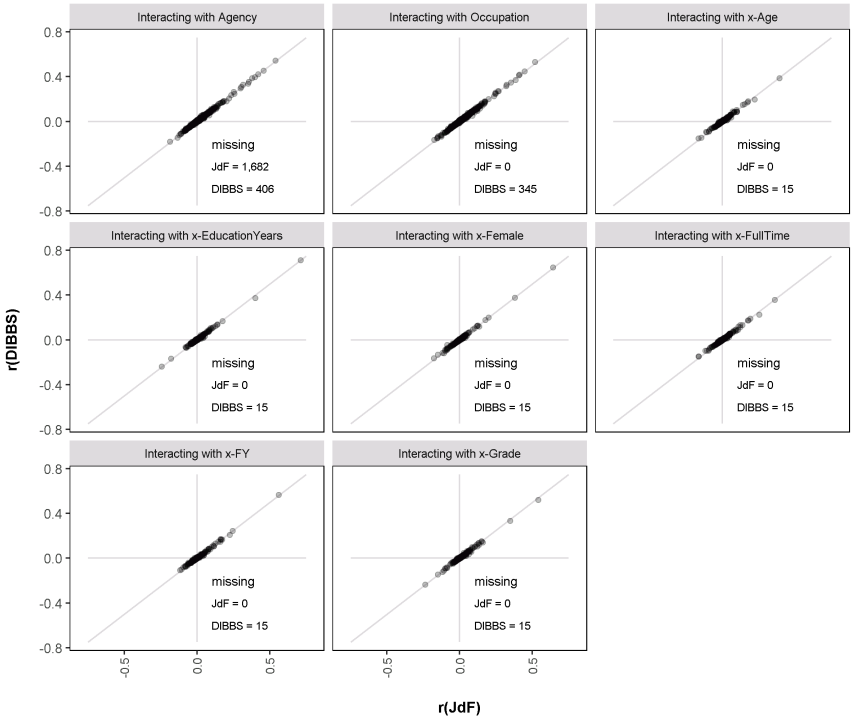
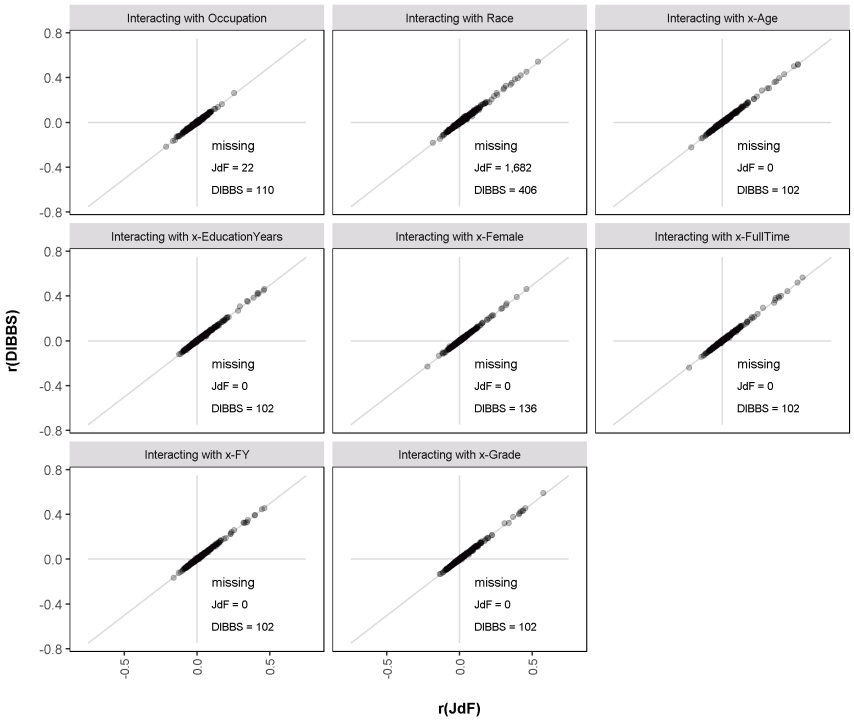


Figure 2: Correlation of primary variables with two variable interactions. Variable set one, involving sex. Synthetic level correlation on y-axis, corresponding authentic correlation on x-axis.

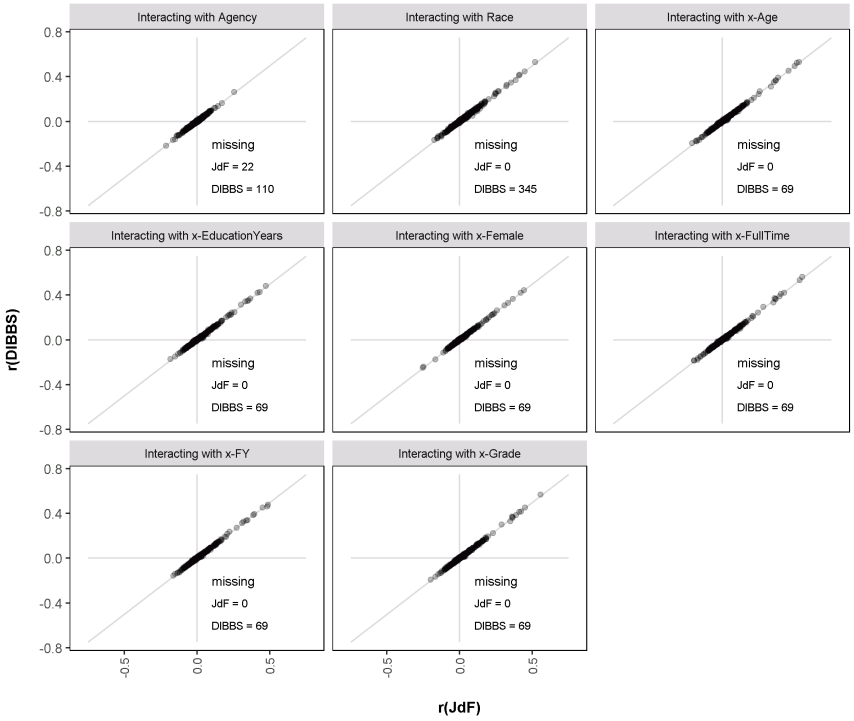


(a) Correlations involving race

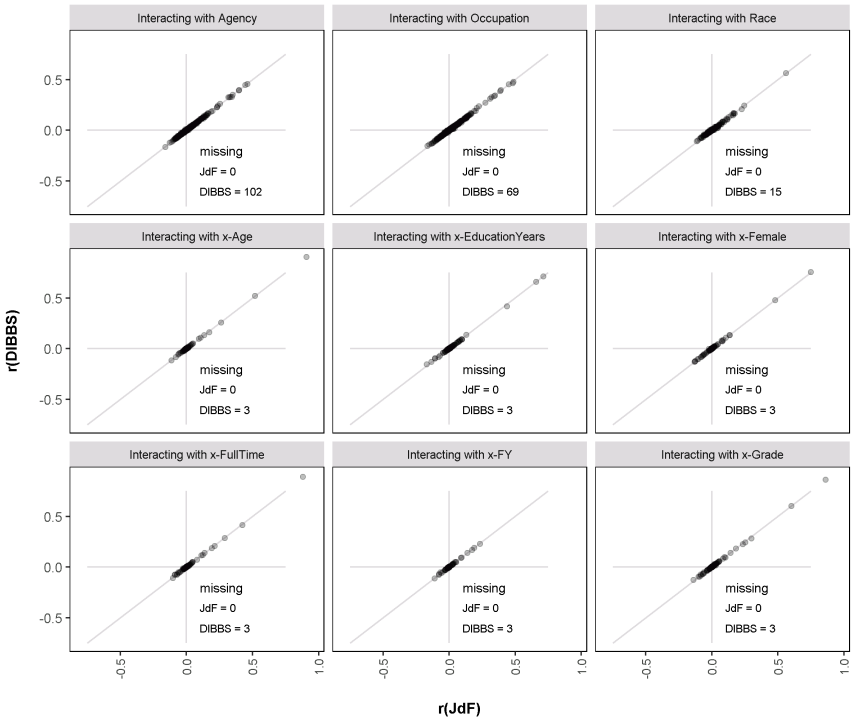


(b) Correlations involving agency

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

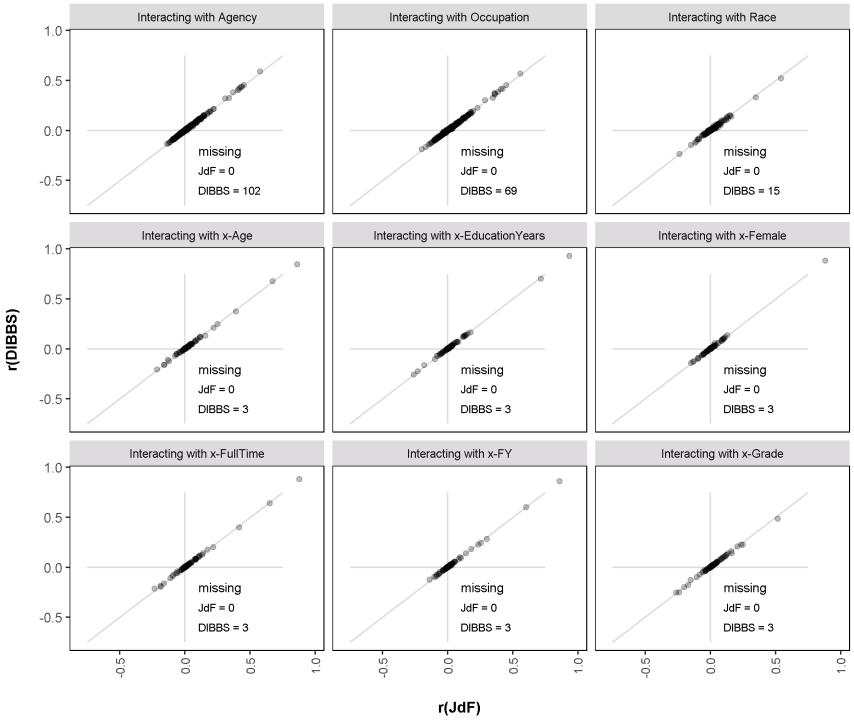


(a) Correlations involving occupation

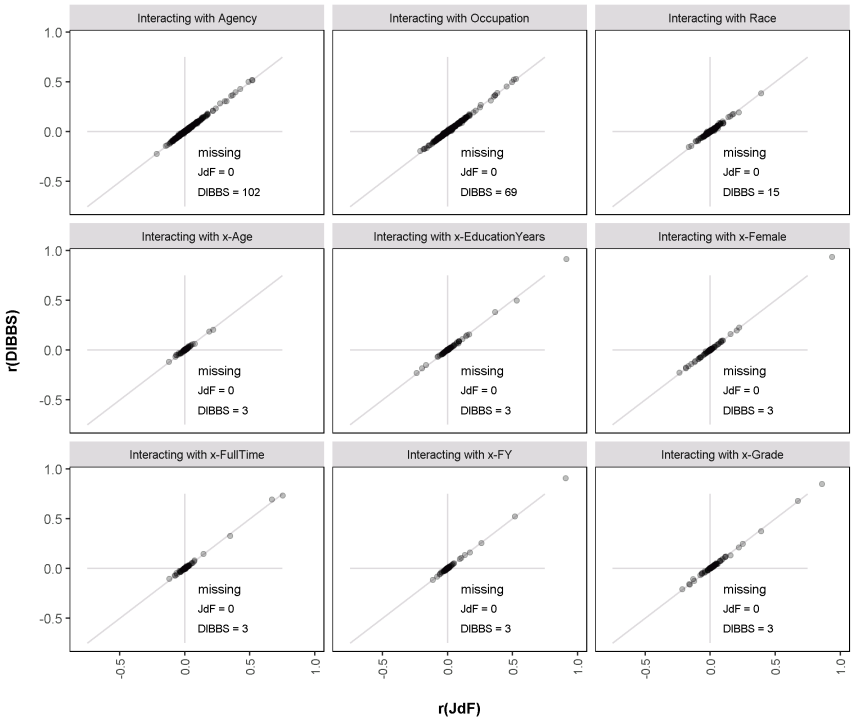


(b) Correlations involving fiscal year

Figure 4: Correlation of primary variables with two variable interactions. Variable set three. Synthetic level correlation on y-axis, corresponding authentic correlation on x-axis.

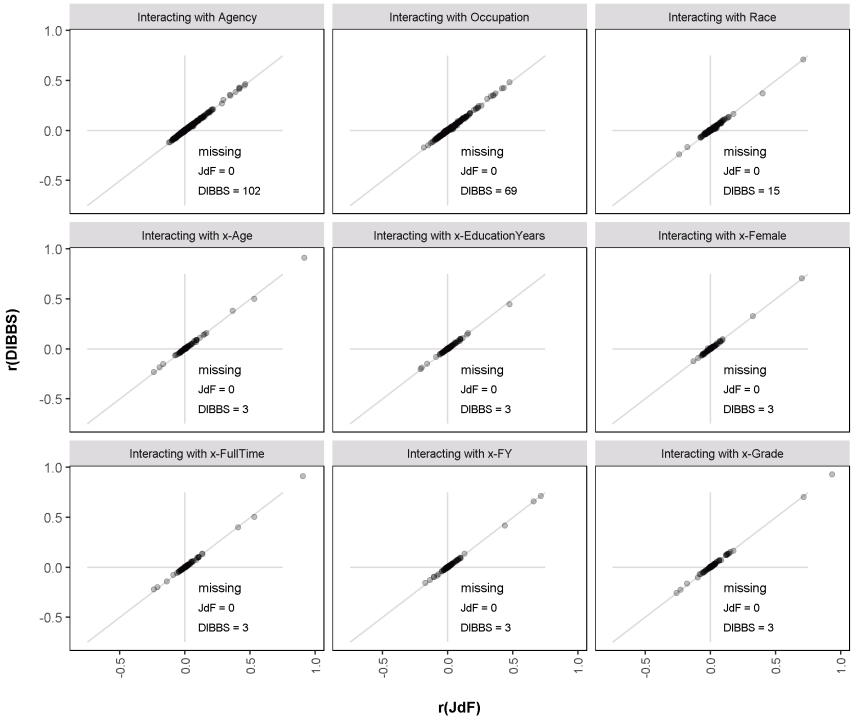


(a) Correlations involving grade

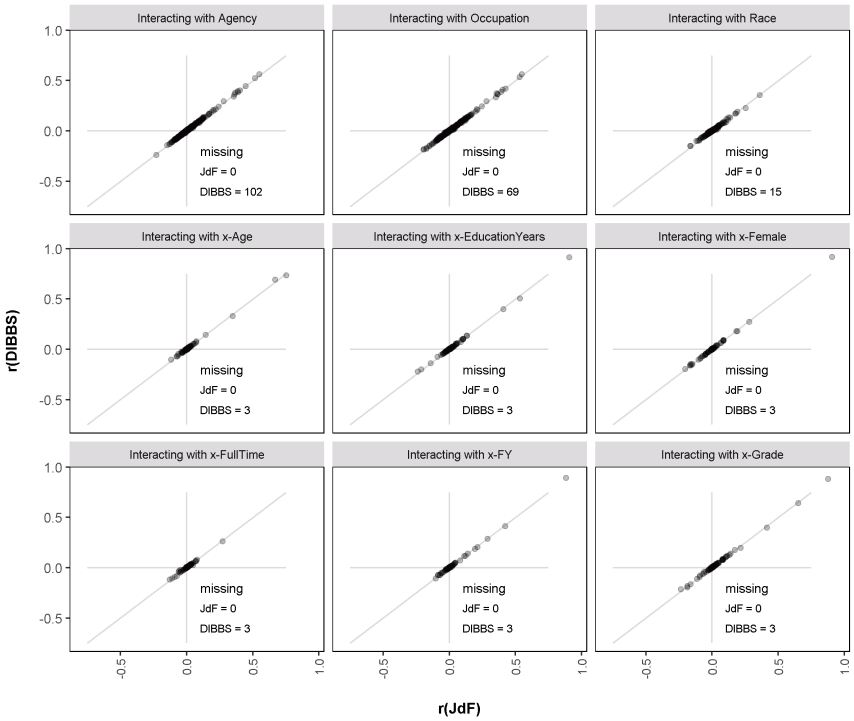


(b) Correlations involving age

Figure 5: Correlation of primary variables with two variable interactions. Variable set four. Synthetic level correlation on y-axis, corresponding authentic correlation on x-axis.



(a) Correlations involving education



(b) Correlations involving work schedule

Figure 6: Correlation of primary variables with two variable interactions. Variable set five. Synthetic level correlation on y-axis, corresponding authentic correlation on x-axis.

3 Cumulative Mass (Proportion Observations) by Pay Plan and Occupation

Pay plan and occupation characterize an employee's job classification and services performed, making them important identifiers in human capital research. Figures 7 and 8 contain example cumulative mass plots for joint combinations of pay plan and occupation. All occupations within each pay plan are represented. Solid line for authentic data, dashed line for synthetic. "nJ" indicates observation count in authentic data, "nD" indicates synthetic data observation count.
Observations: Overlapping or nearness of lines indicates equality of cumulative mass for corresponding levels of occupation within pay plan. 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.

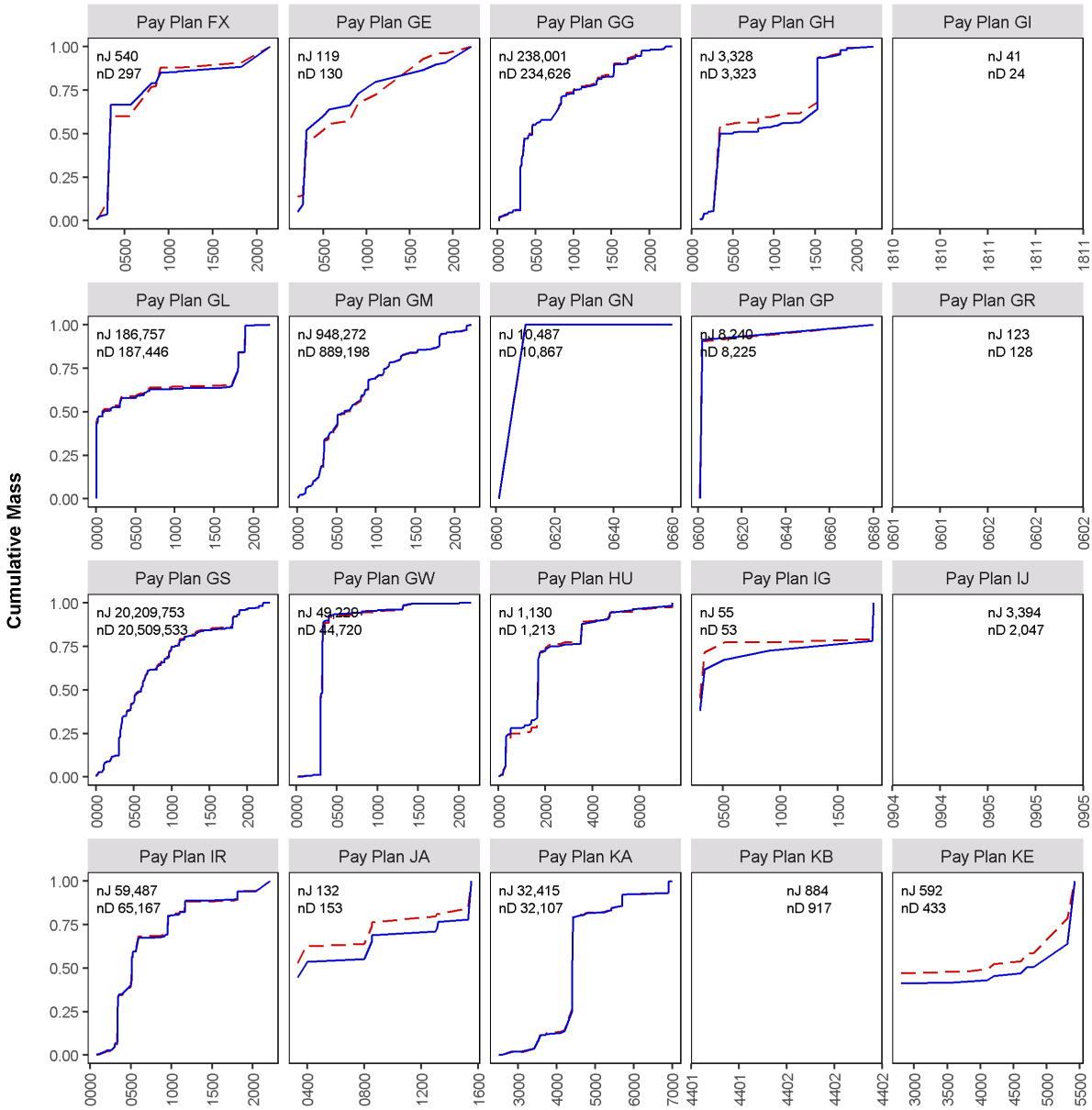


Figure 7: Cumulative mass (observations) by occupation within pay plan. Pay plan set one. Synthetic data represented by dashed line, authentic by solid line.

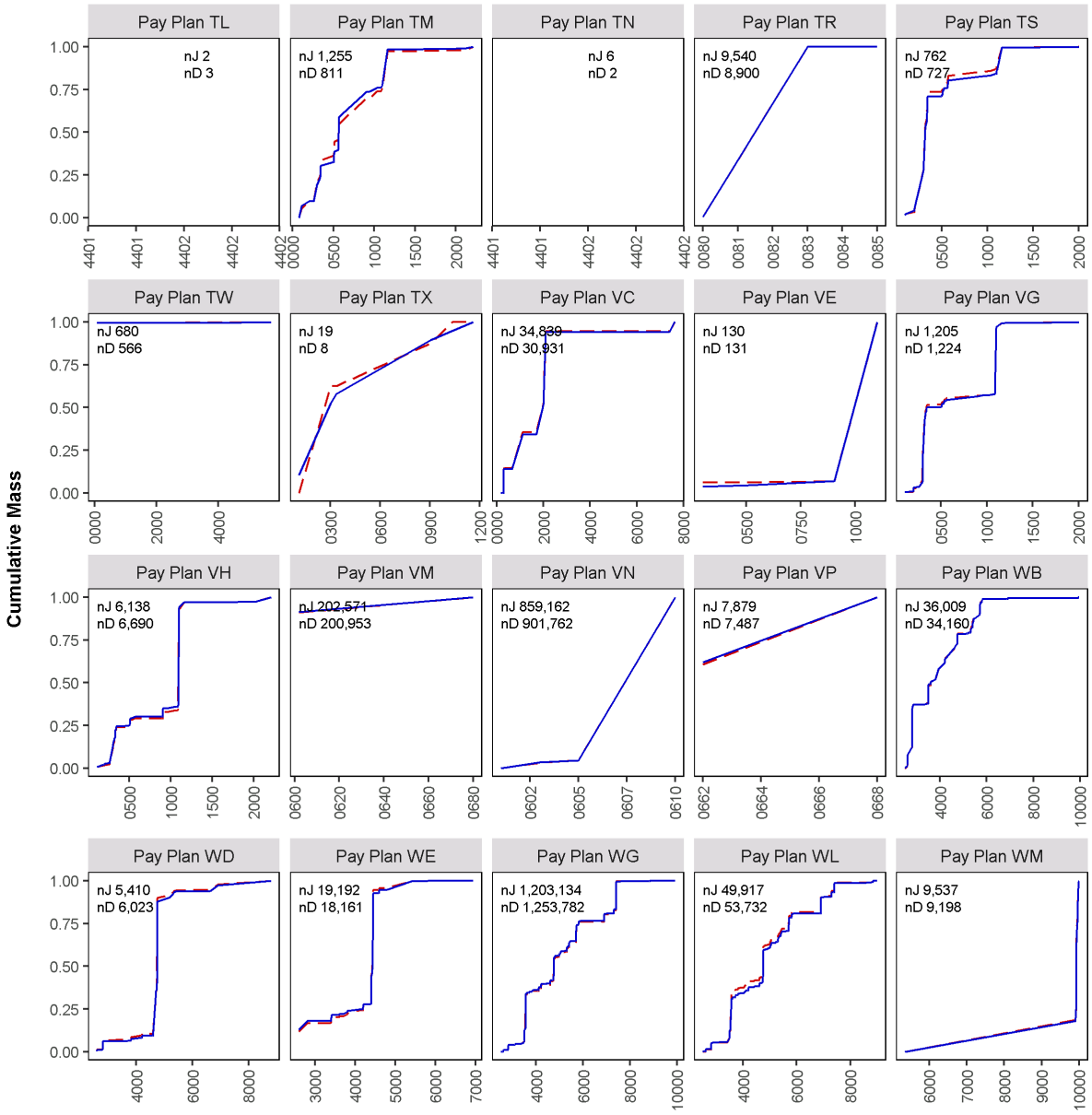


Figure 8: Cumulative mass (observations) by occupation within pay plan. Pay plan set 2. Synthetic data represented by dashed line, authentic by solid line.

4 Distribution of Grade by Joint Combinations of Agency, Occupation, Education, and Supervisor Status

Federal employee pay classification consists of a pay plan, grade, and step rate. Since employee pay is largely determined by grade and promotions are detected by change in grade, it is important for the synthetic data to accurately reflect authentic proportions of observations by grade within joint combinations of critical human capital variables.[3] Observations are limited to GS pay plan, full-time.
Figure 9 compares synthetic and authentic observation frequencies within joint categories of agency, occupational category, and grade. Administrative and professional occupational categories are combined and appear in the lower plot of the panel and remaining GS categories (technical, clerical, and other white collar) appear as a group in the upper plot. Each agency (x-axis) has two columns of dots, one on the left for authentic data and one on the right for synthetic data. The size of each dot is scaled to represent the number of observations in the corresponding grade listed on the y-axis. Dot sizes range from small (less than 1,000 observations) to large (more than 250,000 observations). Total observation counts by agency are listed in rows at the top of each plot, "n(JdF)" indicating authentic observations, "n(DIBBS)" indicating synthetic. Note that the graphs shown are for a sequence of representative agencies. Those for remaining agencies show similar patterns.

Figure 10 compares synthetic and authentic observation frequencies within joint categories of agency, education level (college or not), and grade. Graphs shown are for a sequence of representative agencies. Those for remaining agencies show similar patterns.

Figure 11 compares synthetic and authentic observation frequencies within joint categories of agency, supervisor status, and grade. Graphs shown are for a sequence of representative agencies. Those for remaining agencies show similar patterns.

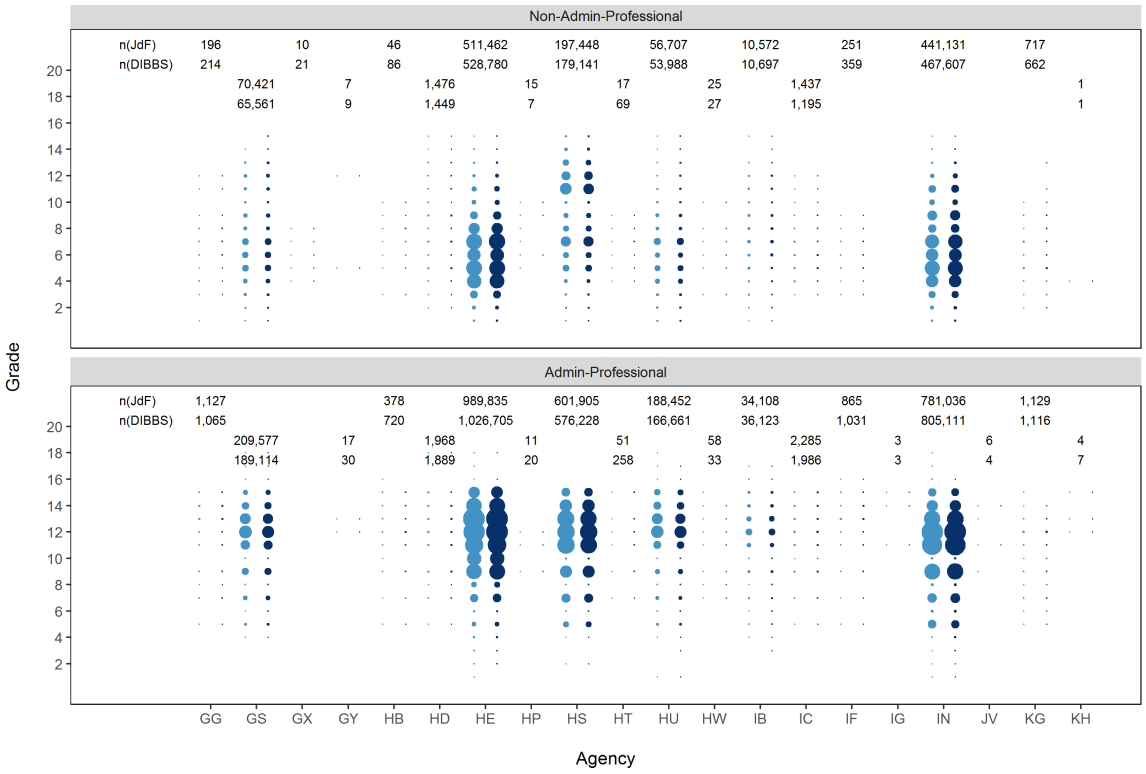
Figure 12 compares synthetic and authentic observation frequencies within joint categories of occupation, occupational category, and grade. Since most occupations are restricted to a single occupational category, panels appear disjoint, showing observations in one category, but not the other. Synthetic data frequencies accurately reflect this property. Graphs shown are for a sequence of representative occupations. Those for remaining occupations show similar patterns.

Figure 13 compares synthetic and authentic observation frequencies within joint categories of occupation, education level (college or not), and grade. Graphs shown are for a sequence of representative occupations. Those for remaining occupations show similar patterns.

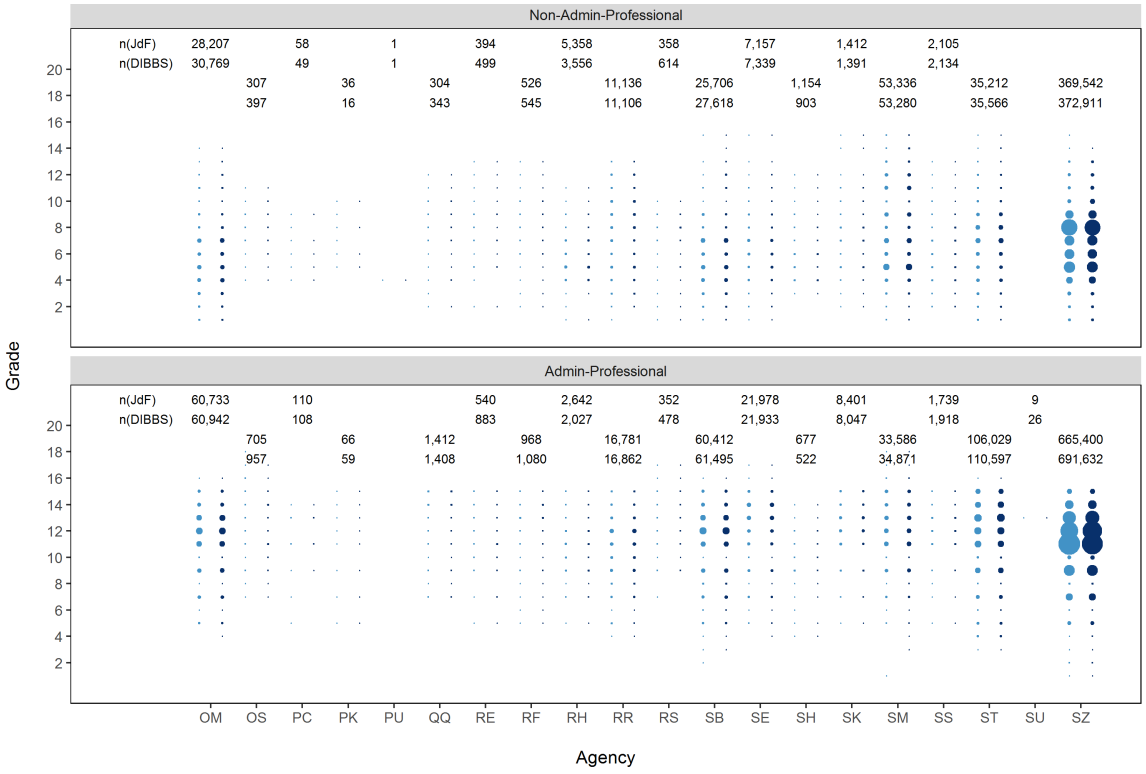
Figure 14 compares synthetic and authentic observation frequencies within joint categories of occupation, supervisor status, and grade. Graphs shown are for a sequence of representative occupations. Those for remaining occupations show similar patterns.

Observations: Within the resolution of dot size, joint frequencies of agency, occupational category, occupation, education, supervisor status, and grade appear to agree in the data sets. To compare n(DIBBS) and n(JdF) observation counts (aggregated counts for all grades within category combinations), the ratio of difference in synthetic and authentic count to authentic count was computed. Figure 15 plots these ratios against corresponding authentic counts (in log scale). Differences tend to be positive (synthetic greater than authentic) and there is a general decrease in error proportion with increase in n.

[3] Although pay rate schedules exist for each pay plan, exceptions to published rates of pay may apply when an employee is transferred into a position that requires a grade adjustment.

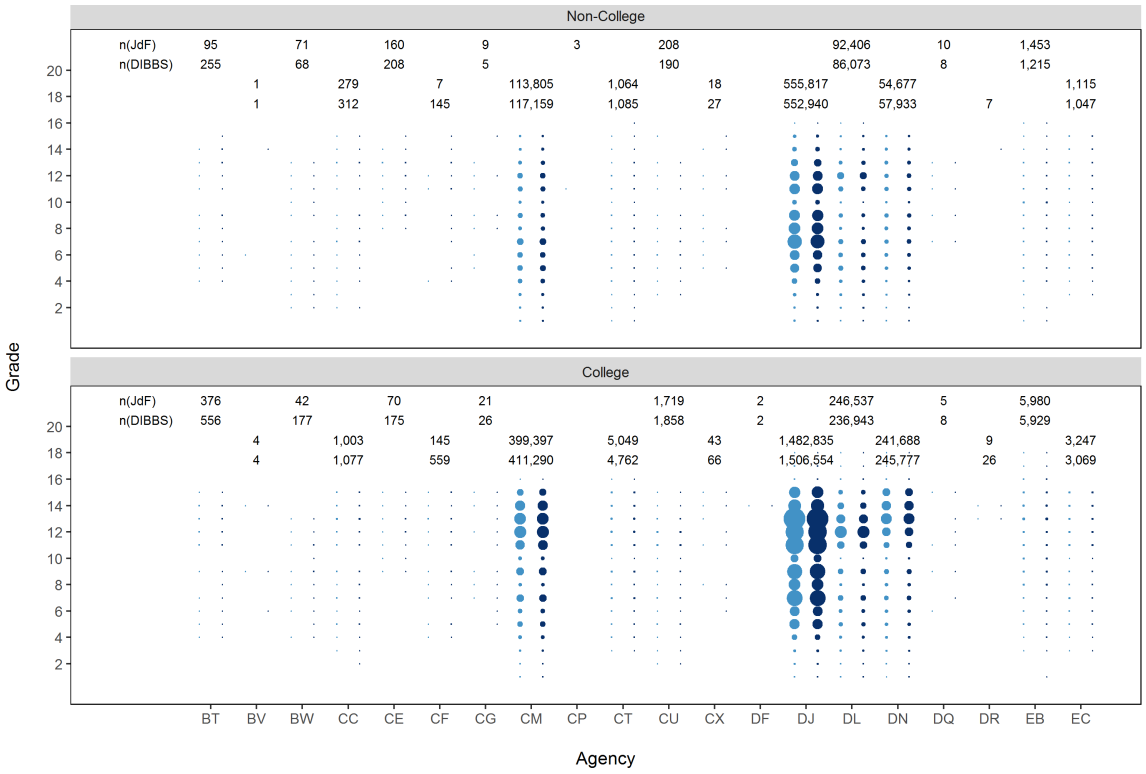


(a) Agencies GG through KH

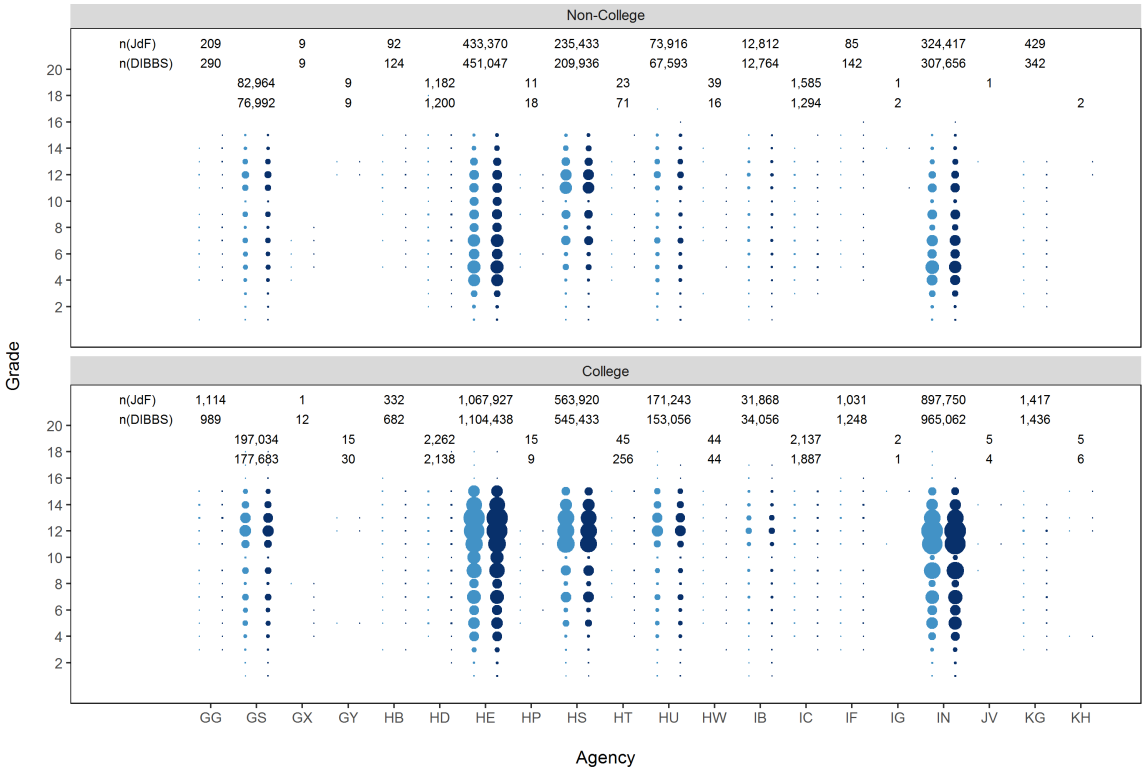


(b) Agencies OM through SZ

Figure 9: Distribution of grade by agency. Occupational categories Admin/Professional vs. non Admin/Professional. Authentic on left, synthetic on right. Dot size small (n¡5,000) to large (n¿250,000).

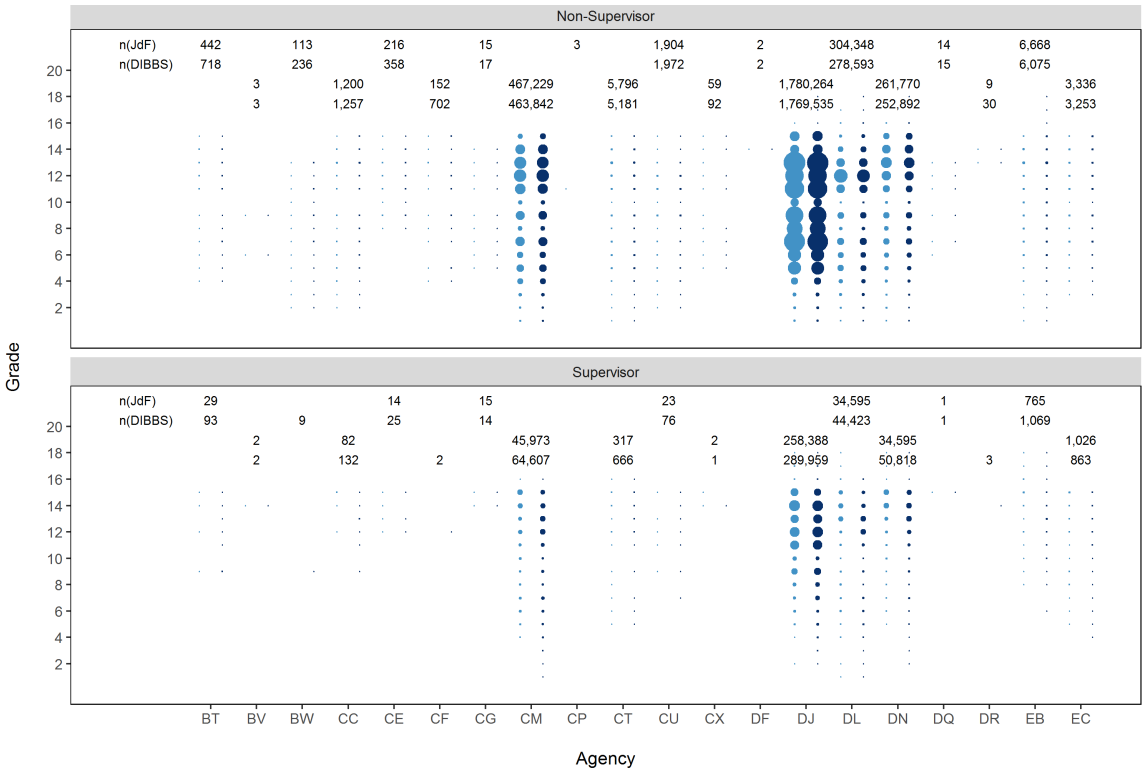


(a) Agencies BT through EC

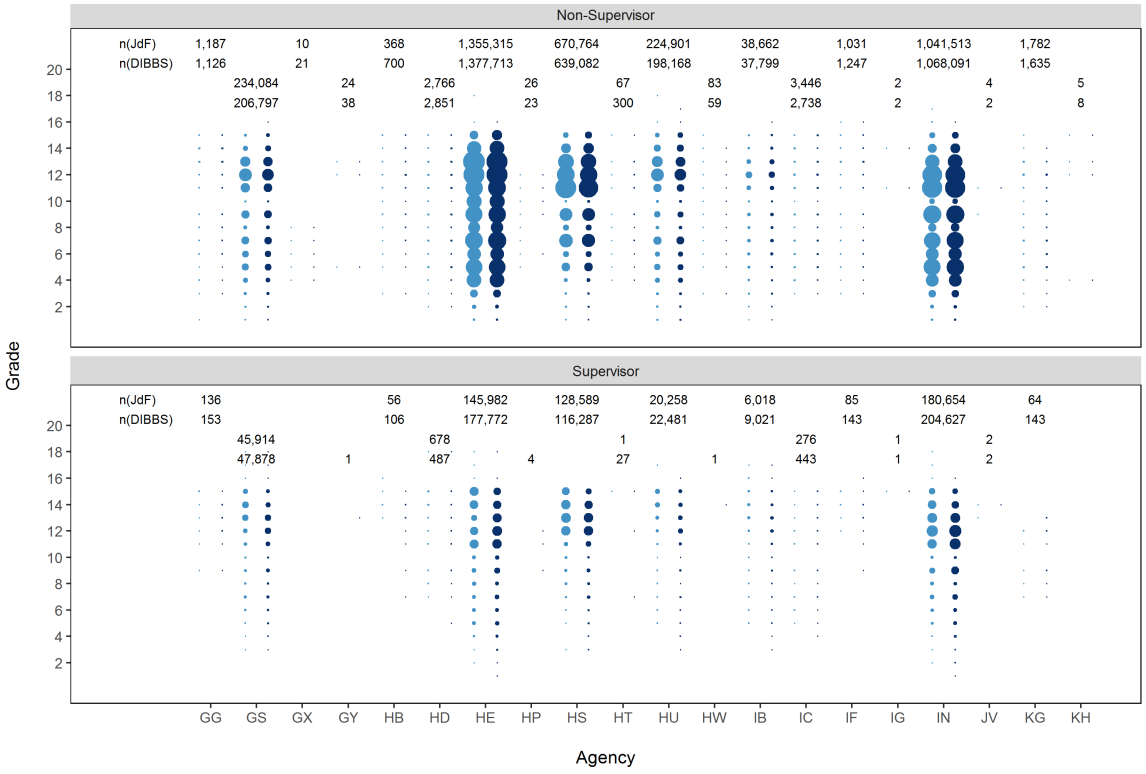


(b) Agencies GG through KH

Figure 10: Distribution of grade by agency. College education vs. non-college. Authentic on left, synthetic on right. Dot size small (n¡5,000) to large (n¿250,000).

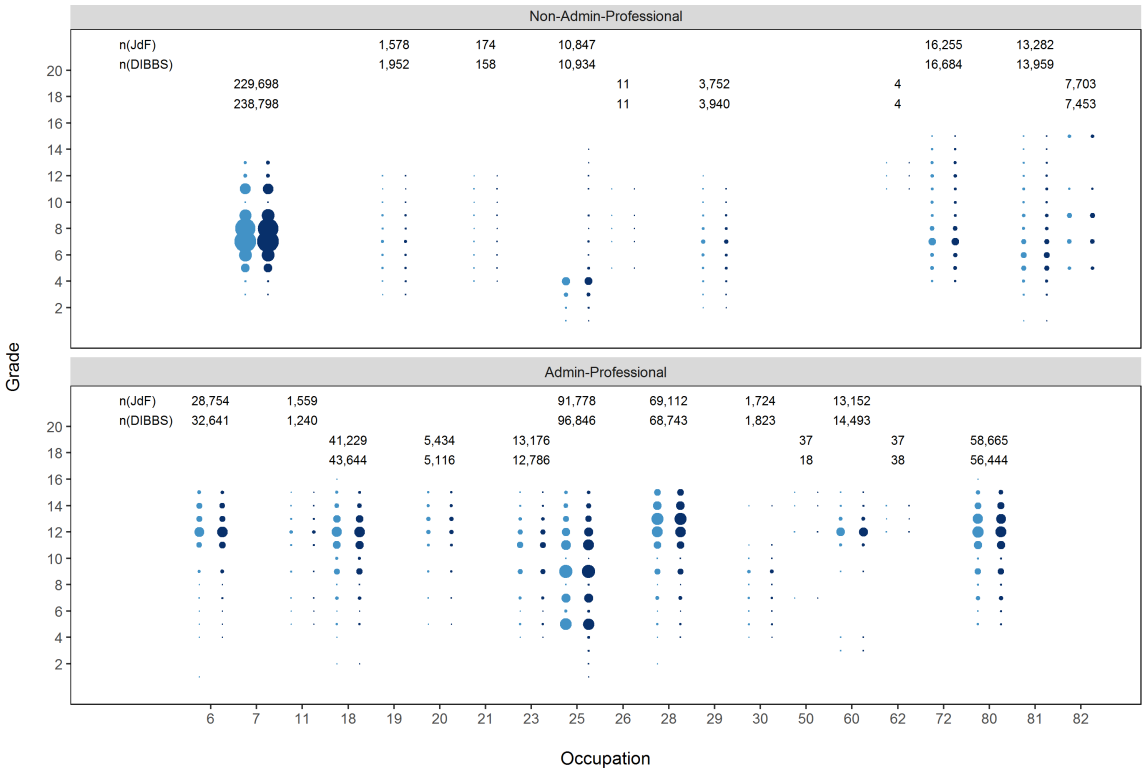


(a) Agencies BT through EC

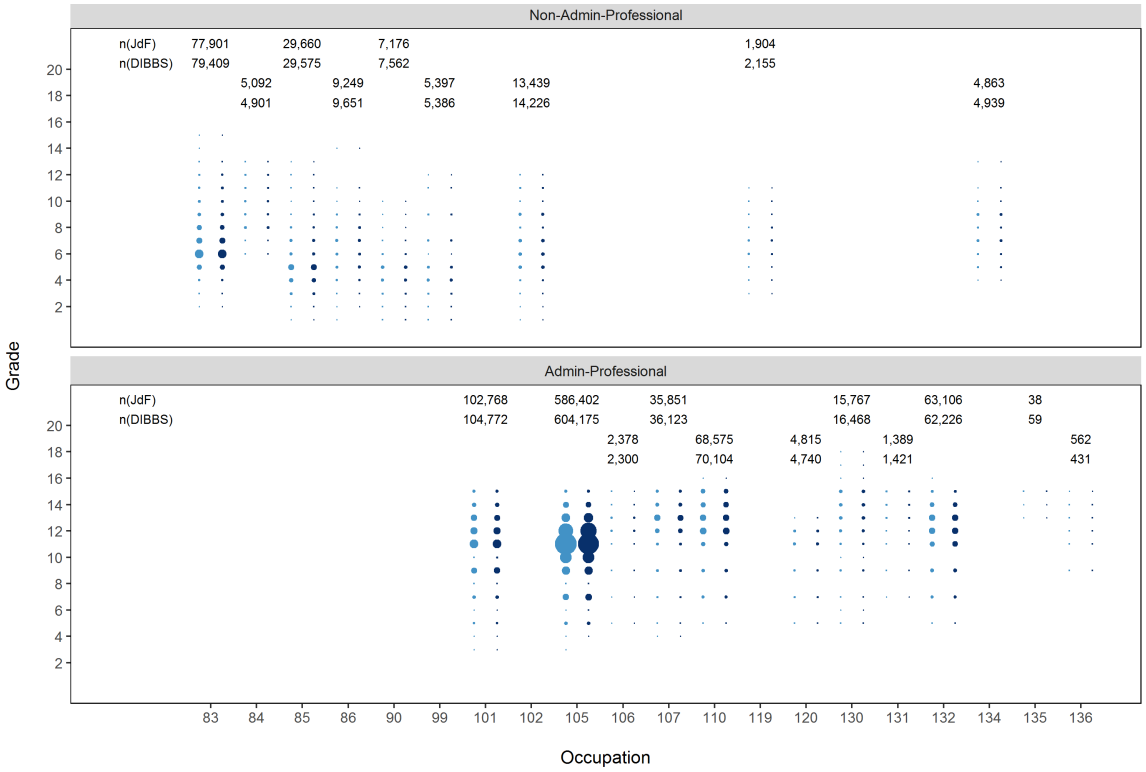


(b) Agencies GG through KH

Figure 11: Distribution of grade by agency and supervisor status. Authentic on left, synthetic on right. Dot size small (n¡5,000) to large (n¿250,000).

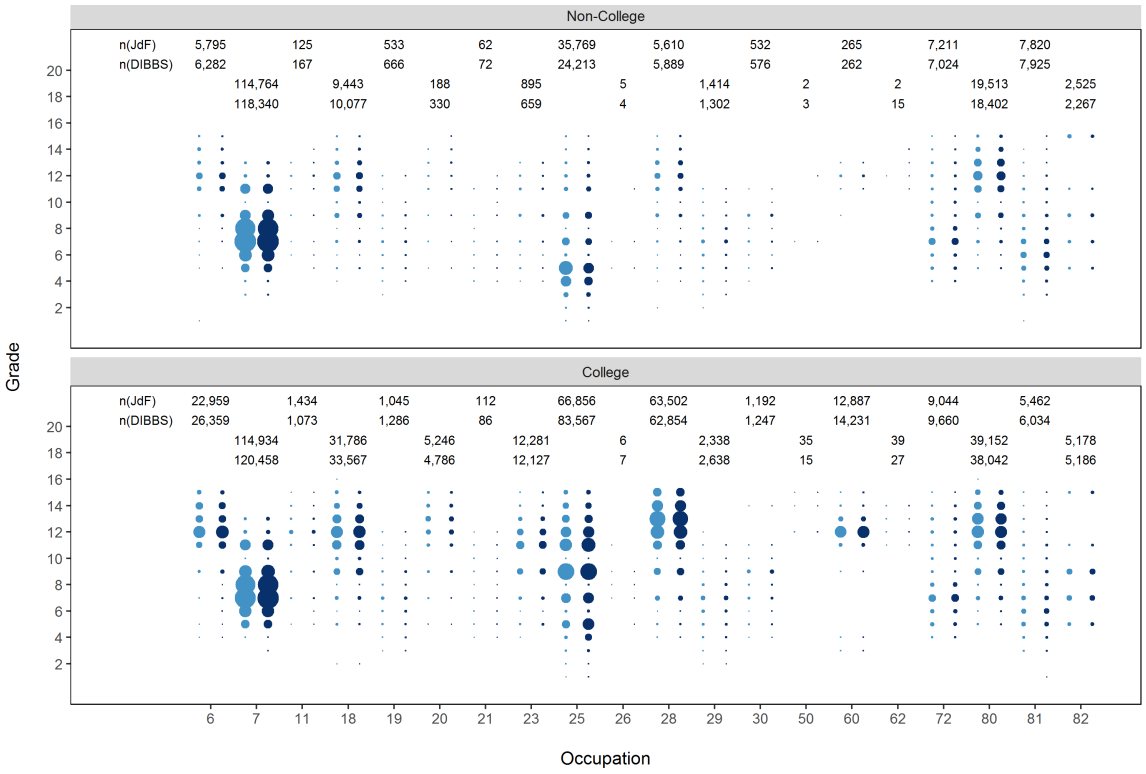


(a) Occupations 0006 through 0082

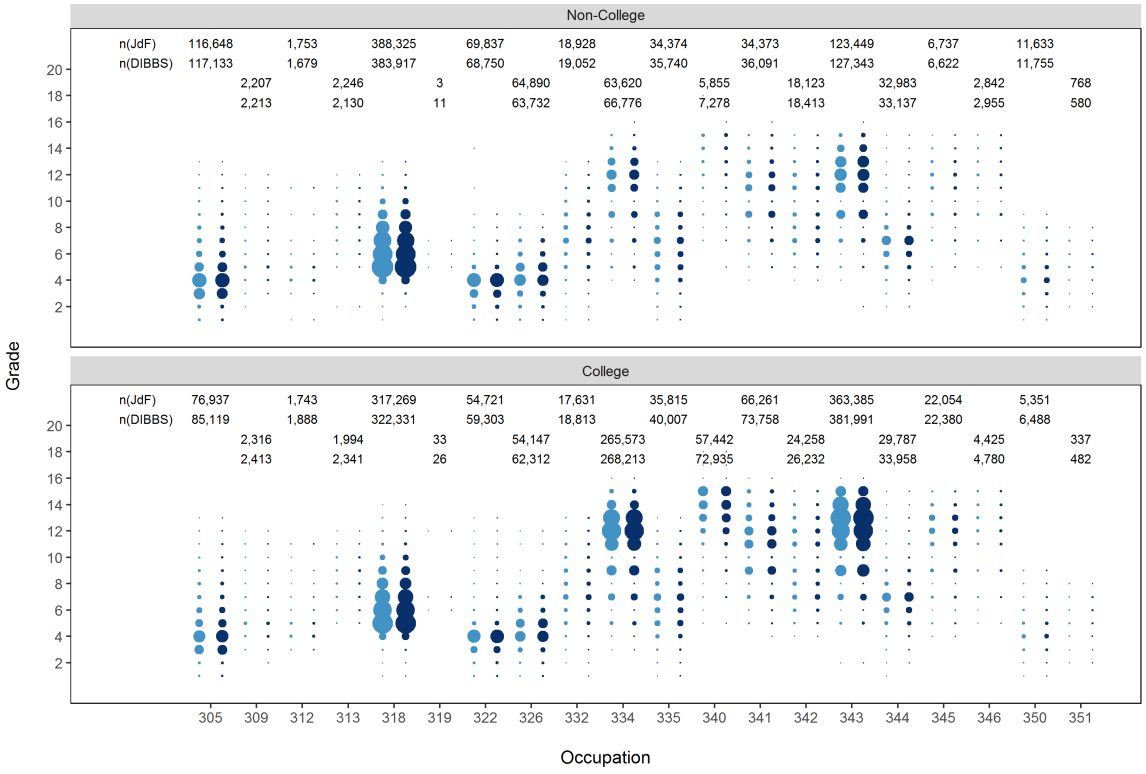


(b) Occupations 0083 through 0136

Figure 12: Distribution of grade by occupation. Occupational categories Admin/Professional vs. non Admin/Professional. Authentic on left, synthetic on right. Dot size small (n¡5,000) to large (n¿250,000).

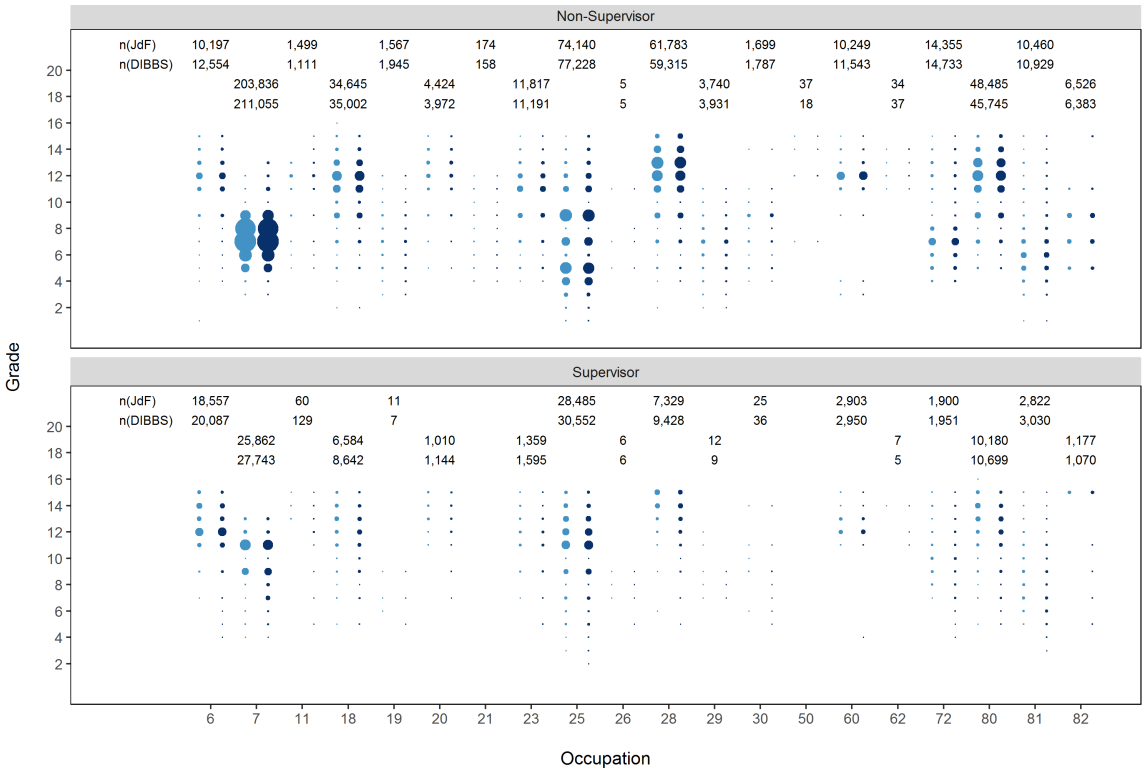


(a) Occupations 0006 through 0082

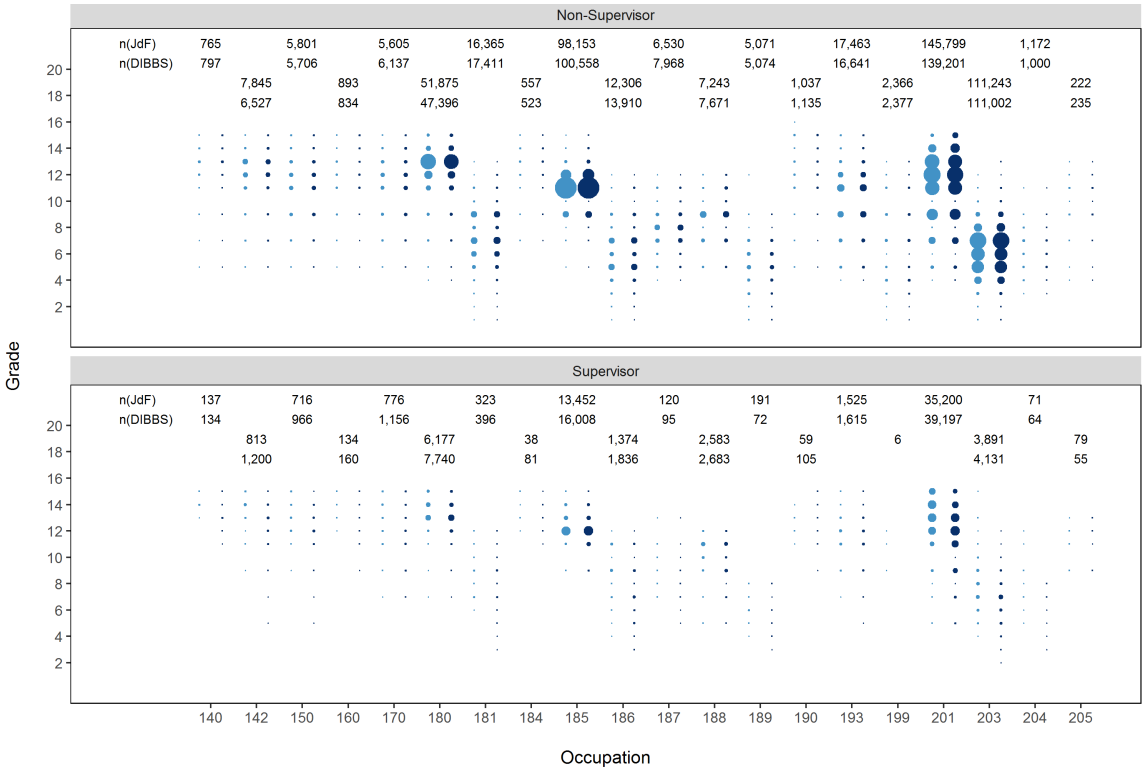


(b) Occupations 0305 through 0351

Figure 13: Distribution of grade by occupation. College education vs. non-college. Authentic on left, synthetic on right. Dot size small (n¡5,000) to large (n¿250,000).



(a) Occupations 0006 through 0082



(b) Occupations 0140 through 0205

Figure 14: Distribution of grade by agency and supervisor status. Authentic on left, synthetic on right. Dot size small (n¡5,000) to large (n¿250,000).

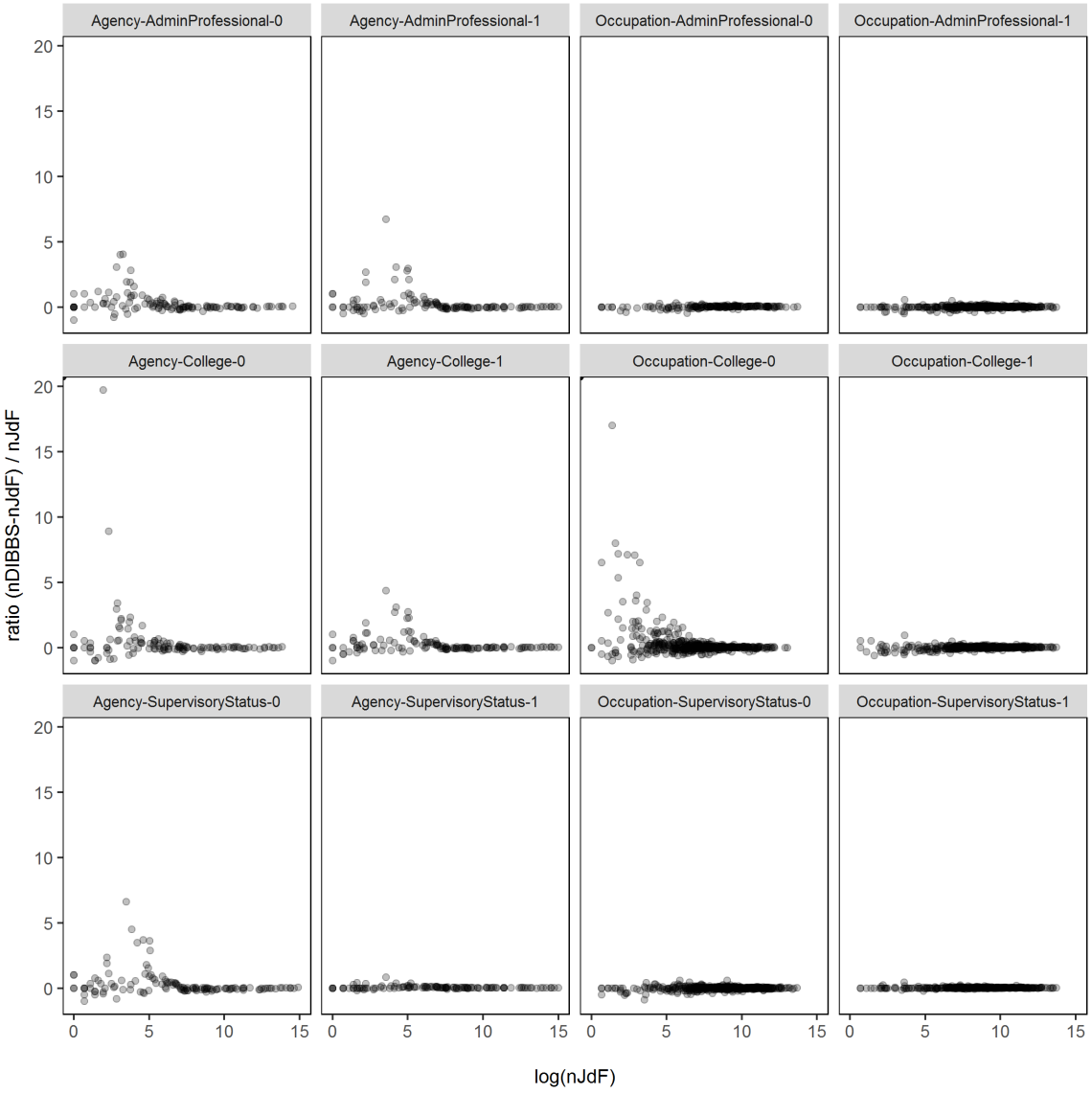


Figure 15: Ratio of difference in synthetic and authentic observation count to authentic observation count. Suffix of 0 indicates non-Admin, non-College, or non-Supervisor, suffix of 1 indicates Admin, College, or Supervisor. Log x-axis scale. Differences tend to be positive (synthetic > authentic). General decrease in error proportion with increase in n.

5 Distribution of Basic Pay

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.[4] [5]

5.1 Distribution of Basic Pay by Agency

Figure 16 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.

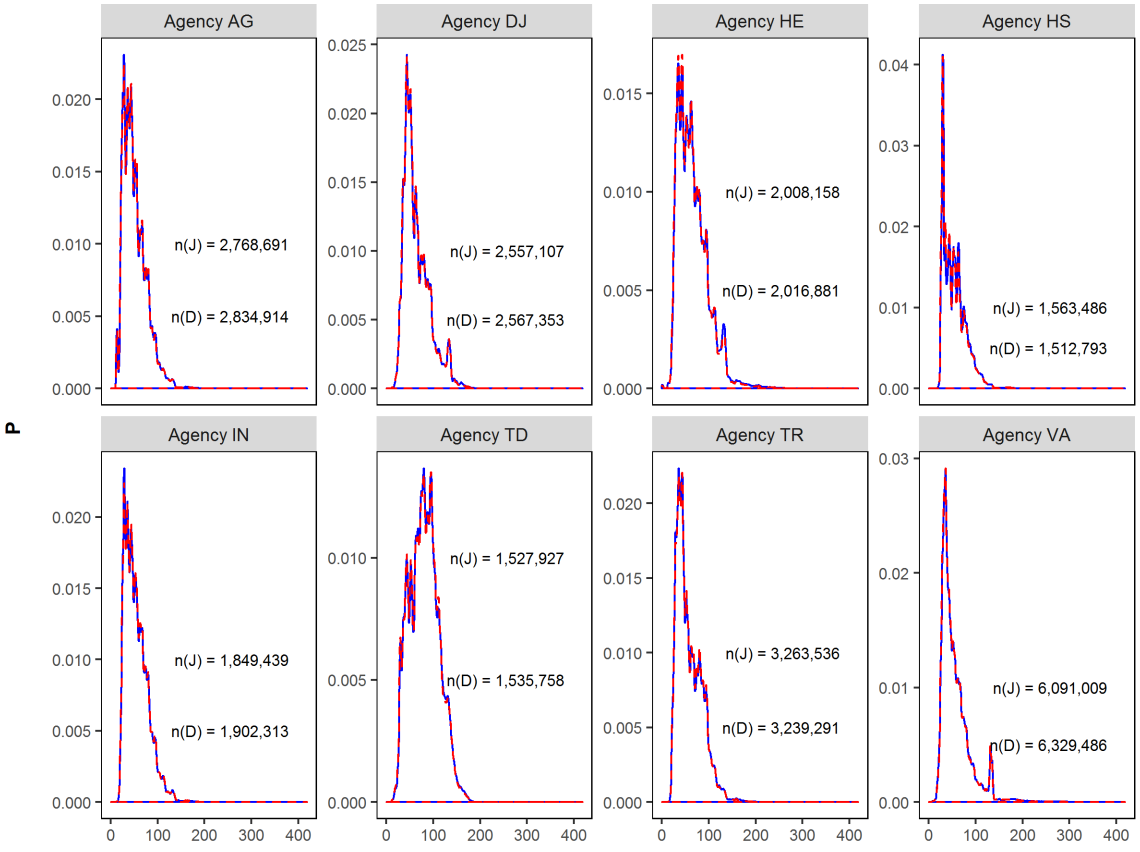


Figure 16: Basic pay marginal distribution for top eight agencies. Dashed line for synthetic data, solid line for authentic.

[4] Basic pay is determined by pay plan, grade, and step rate and does not include adjustments such as for locale.

[5] Due to inflation, the value of wages earned is variable, typically decreasing by year. To compensate for this, and to make inter-year values comparable, pay values throughout the entire document are adjusted to 2011 dollars using annual consumer price indices.

5.2 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 17 through 24 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, supervisory, 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.

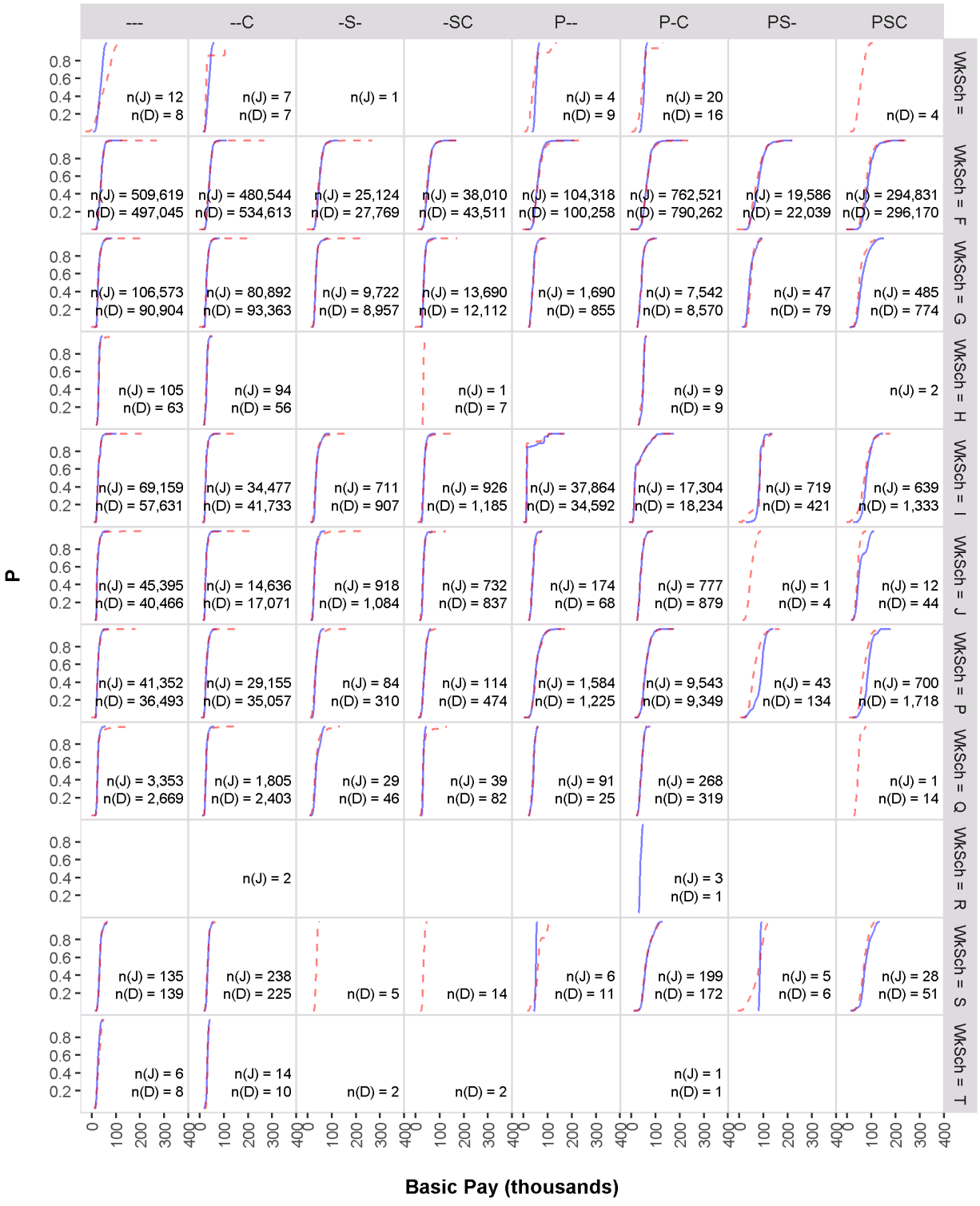


Figure 17: Basic pay distribution by joint professional, supervisory status, college education, and work schedule categories. Department of Agriculture (AG). Dashed line for synthetic data, solid line for authentic.

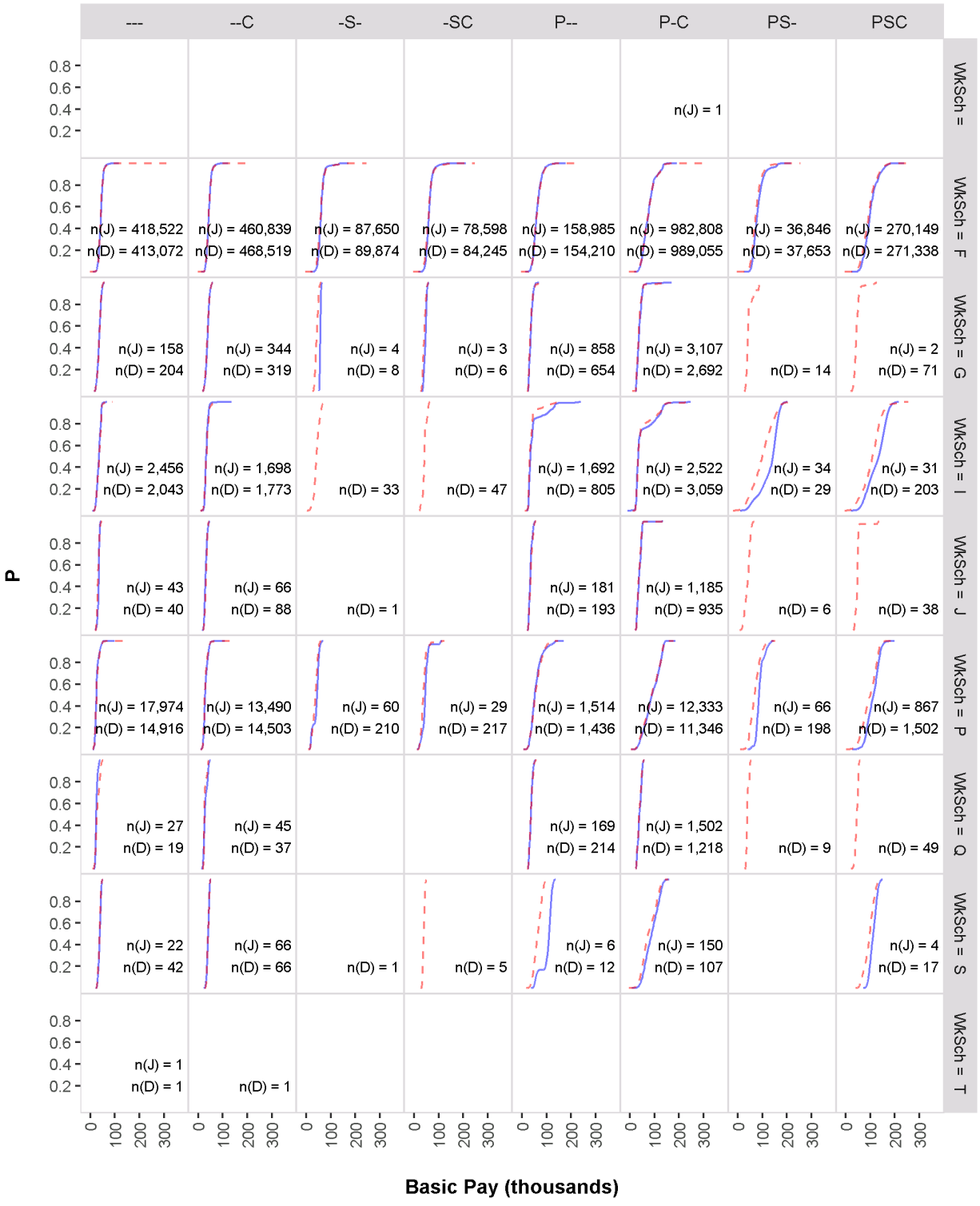


Figure 18: Basic pay distribution by joint professional, supervisory status, college education, and work schedule categories. Department of Justice (DJ). Dashed line for synthetic data, solid line for authentic.

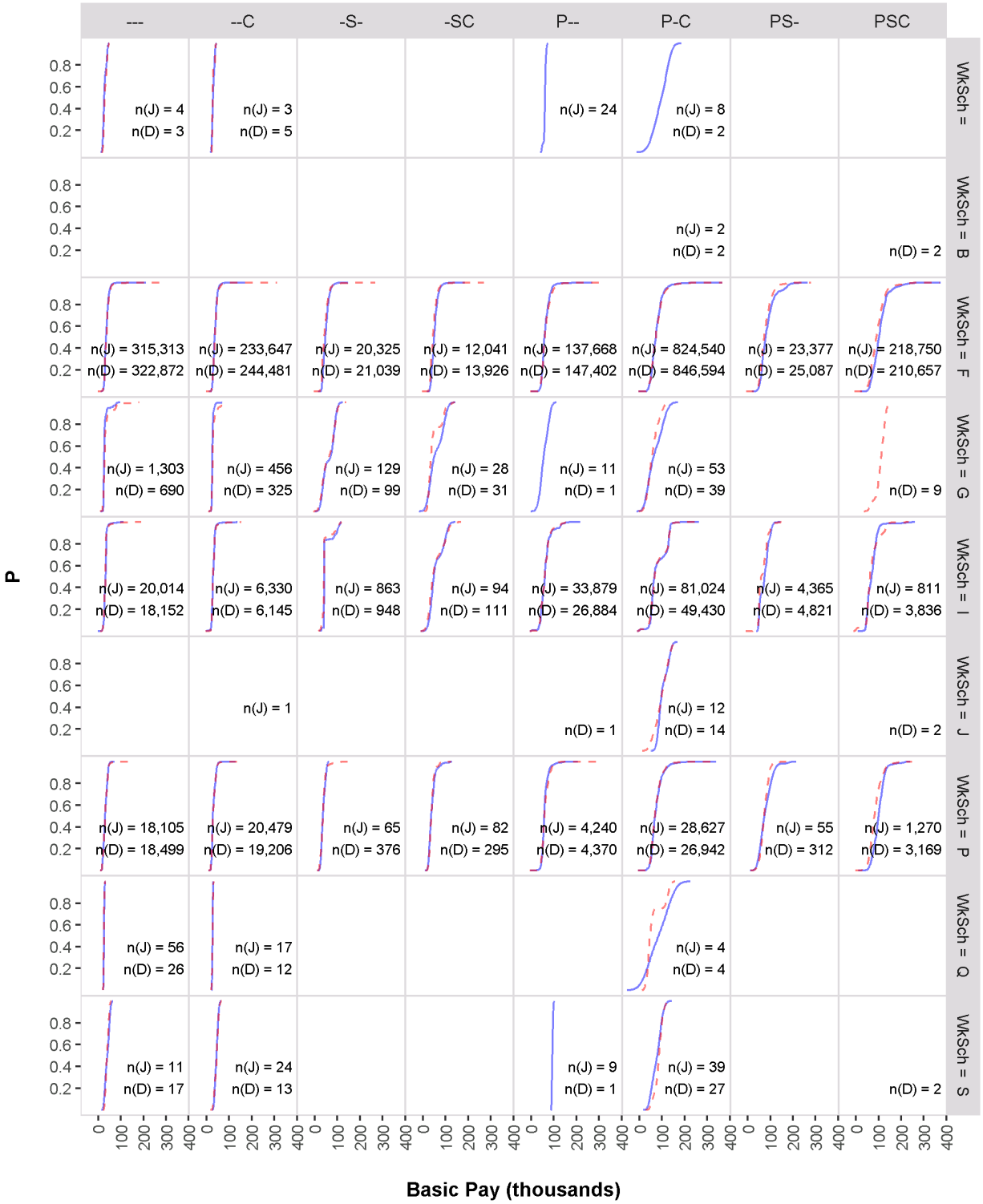


Figure 19: Basic pay distribution by joint professional, supervisory status, college education, and work schedule categories. Department of Health and Human Services (HE). Dashed line for synthetic data, solid line for authentic.

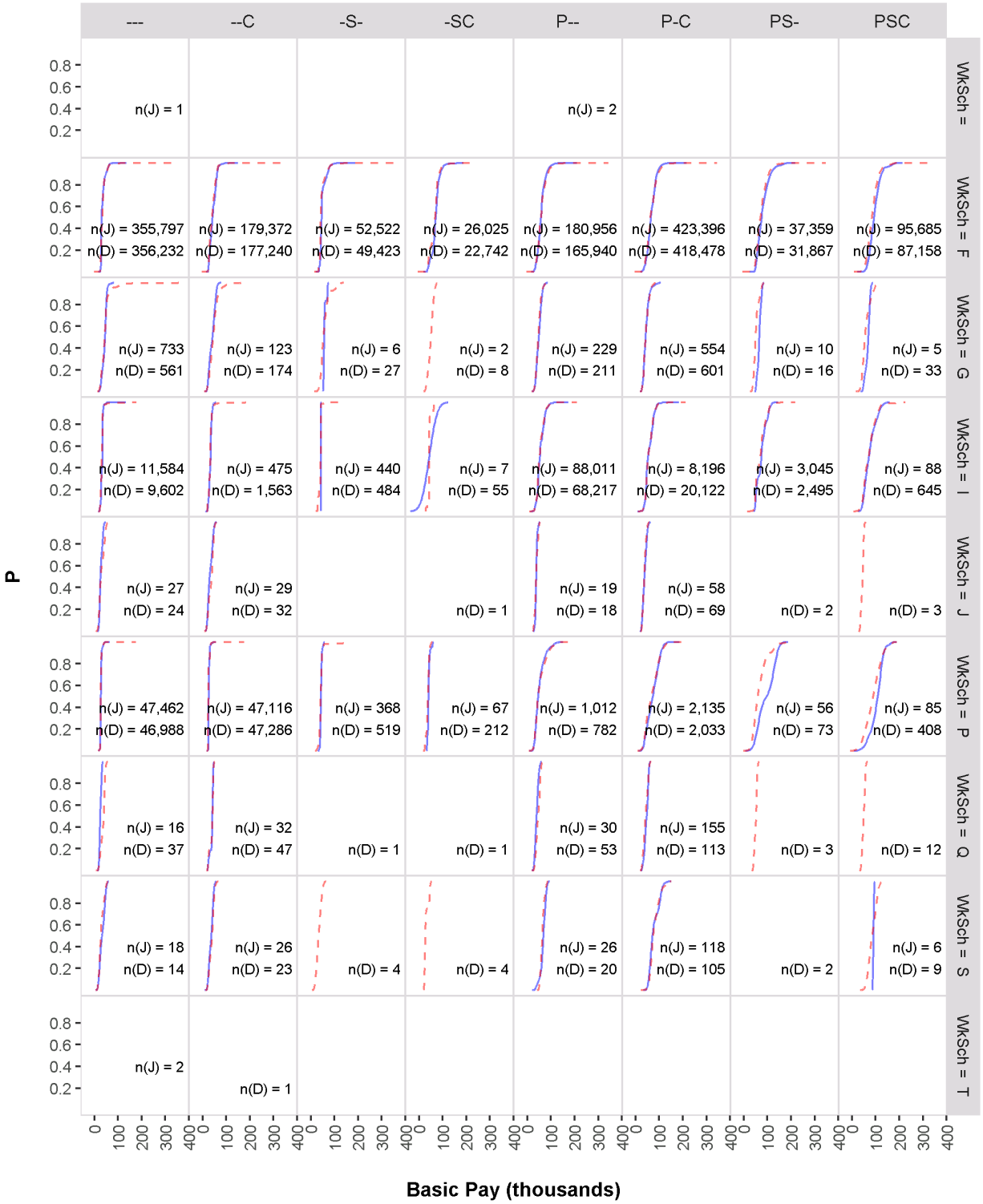


Figure 20: Basic pay distribution by joint professional, supervisory status, college education, and work schedule categories. Department of Homeland Security (HS). Dashed line for synthetic data, solid line for authentic.

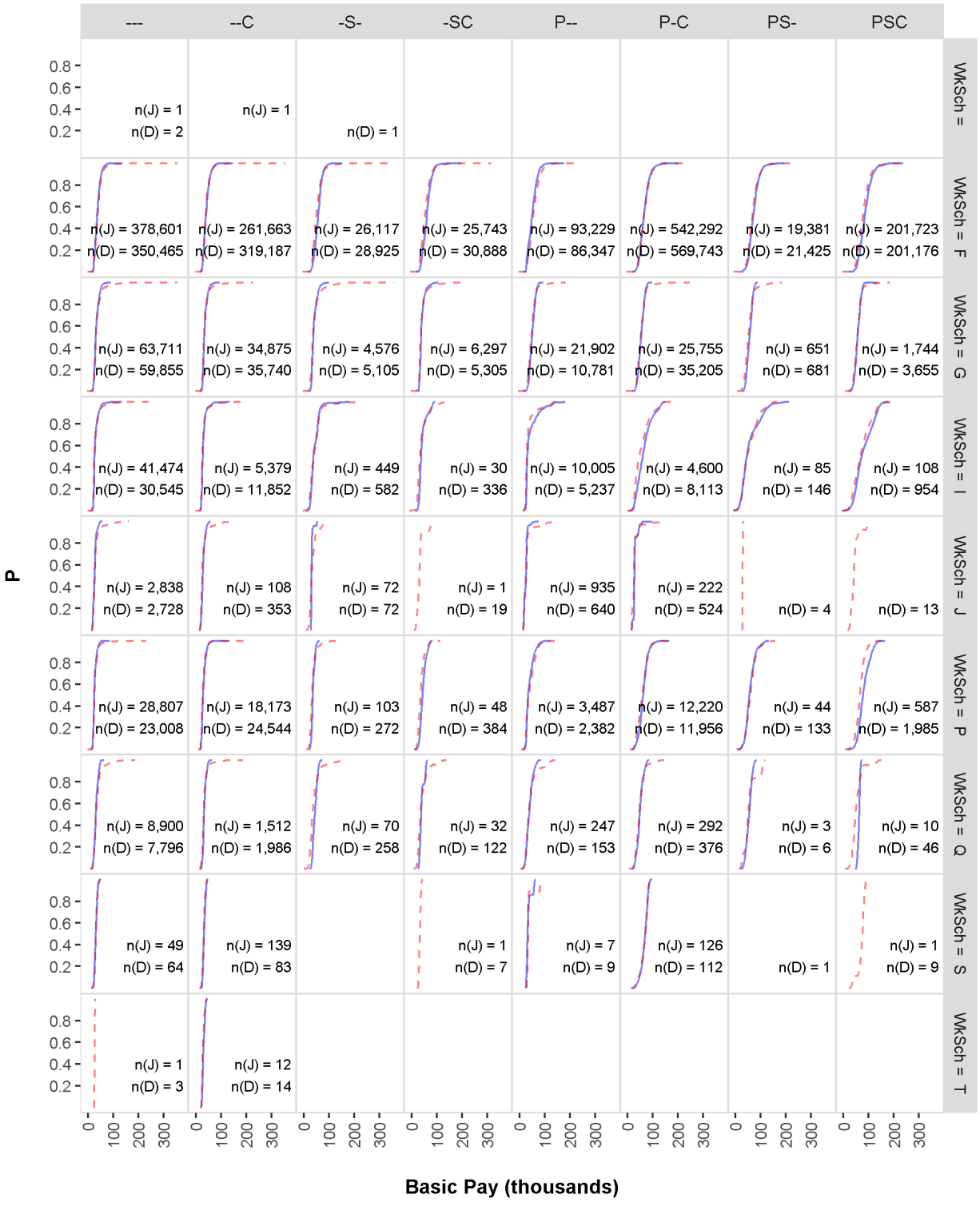


Figure 21: Basic pay distribution by joint professional, supervisory status, college education, and work schedule categories. Department of Interior (IN). Dashed line for synthetic data, solid line for authentic.

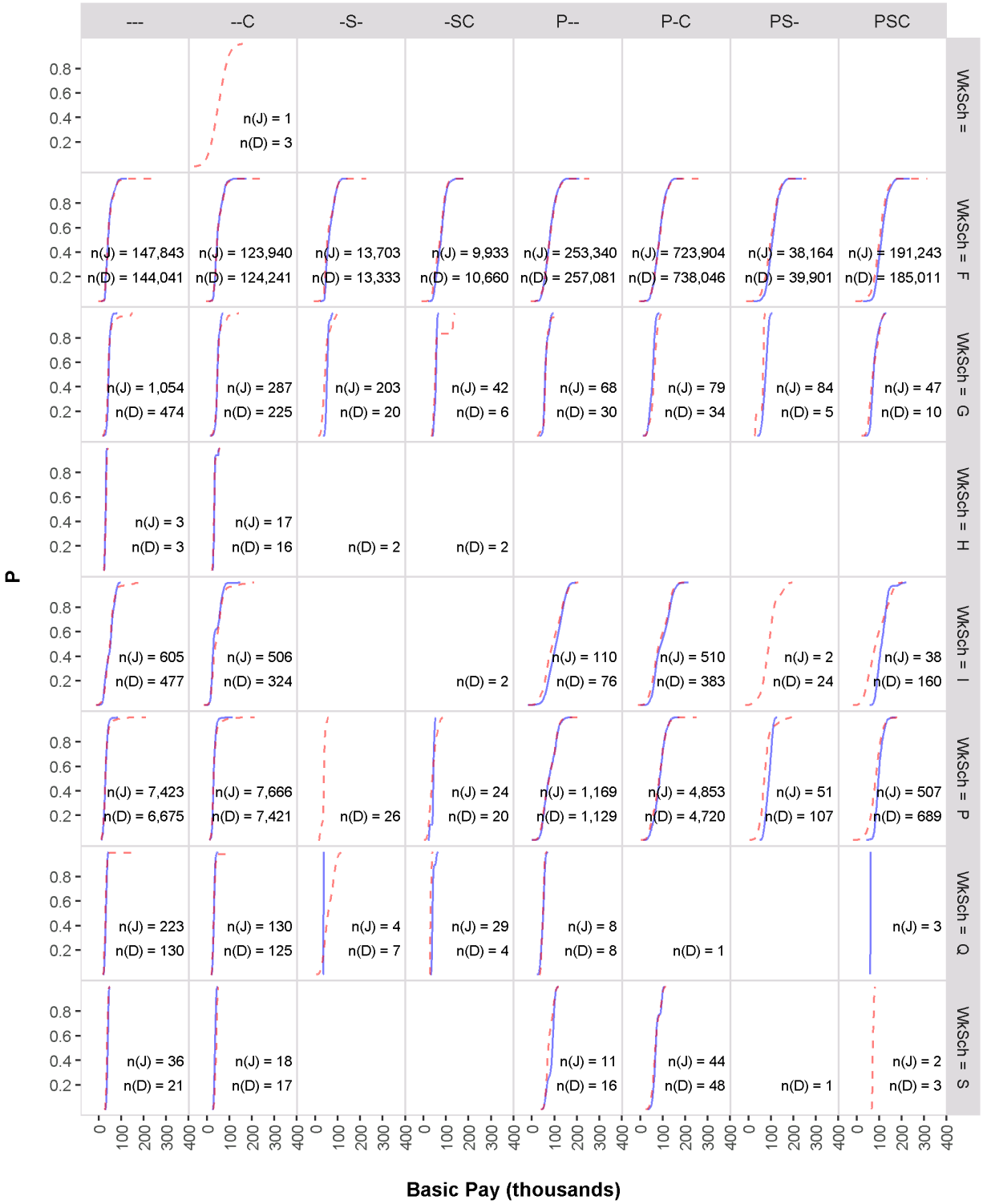


Figure 22: Basic pay distribution by joint professional, supervisory status, college education, and work schedule categories. Department of Transportation (TD). Dashed line for synthetic data, solid line for authentic.

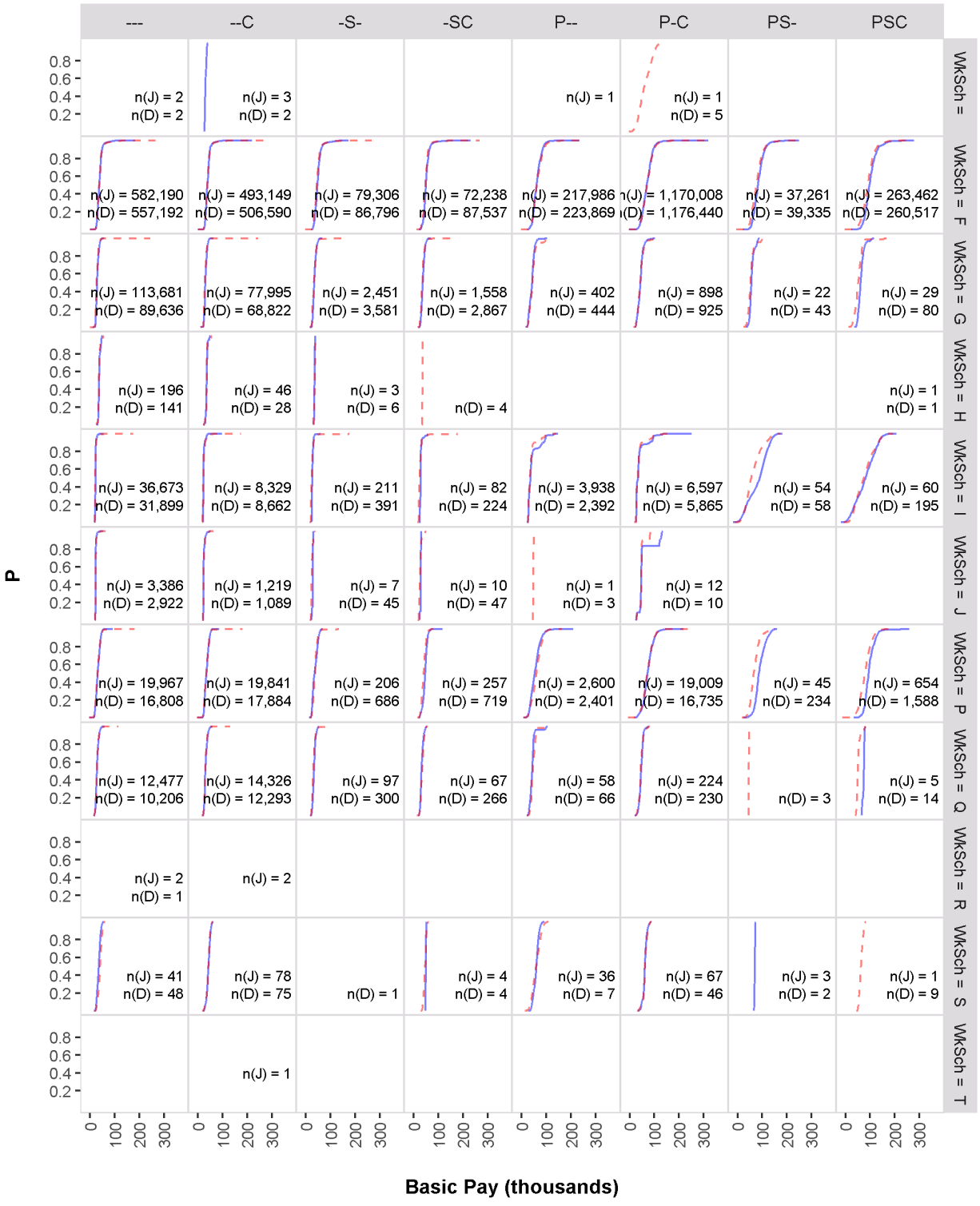


Figure 23: Basic pay distribution by joint professional, supervisory status, college education, and work schedule categories. Department of Treasury (TR). Dashed line for synthetic data, solid line for authentic.

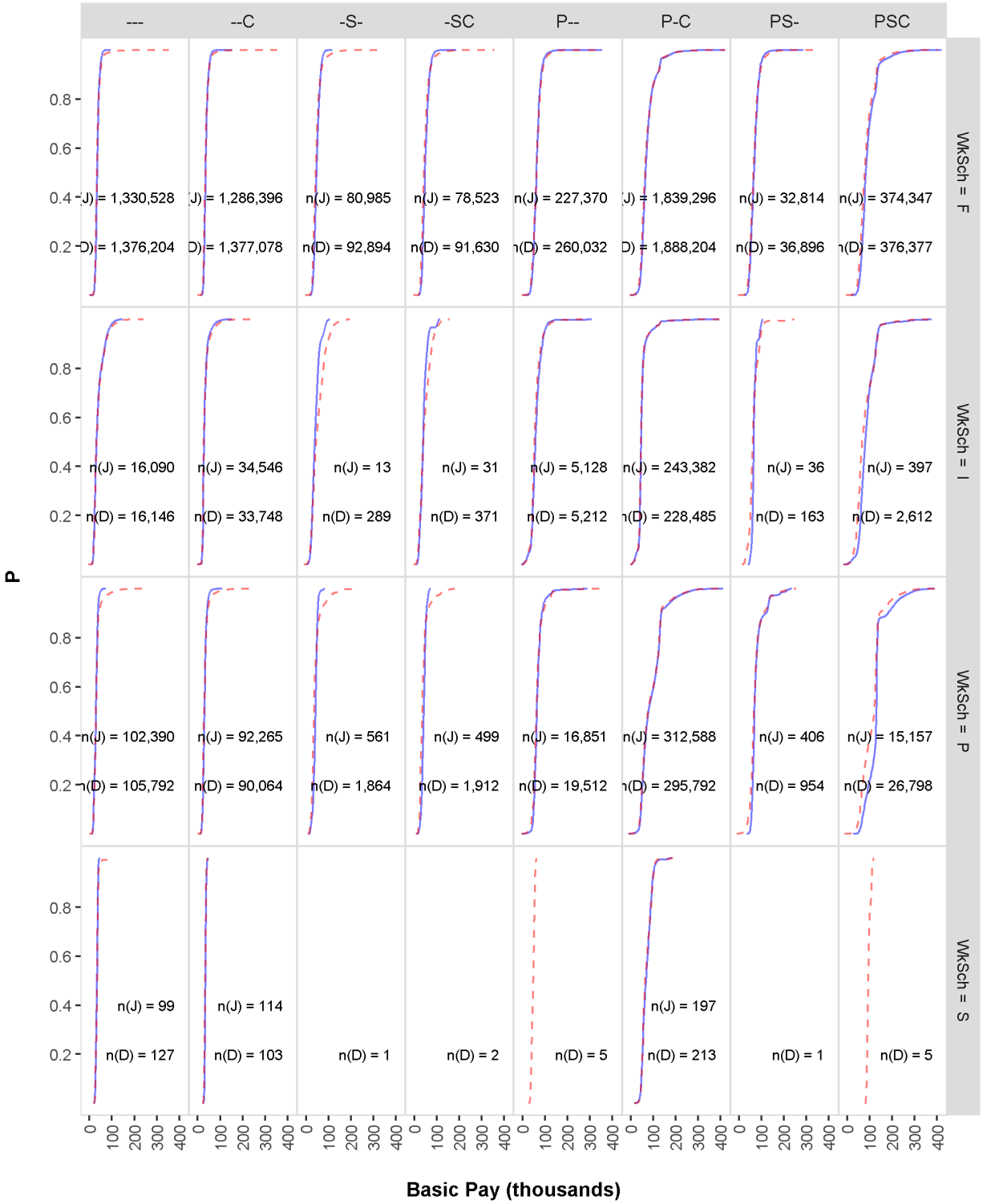
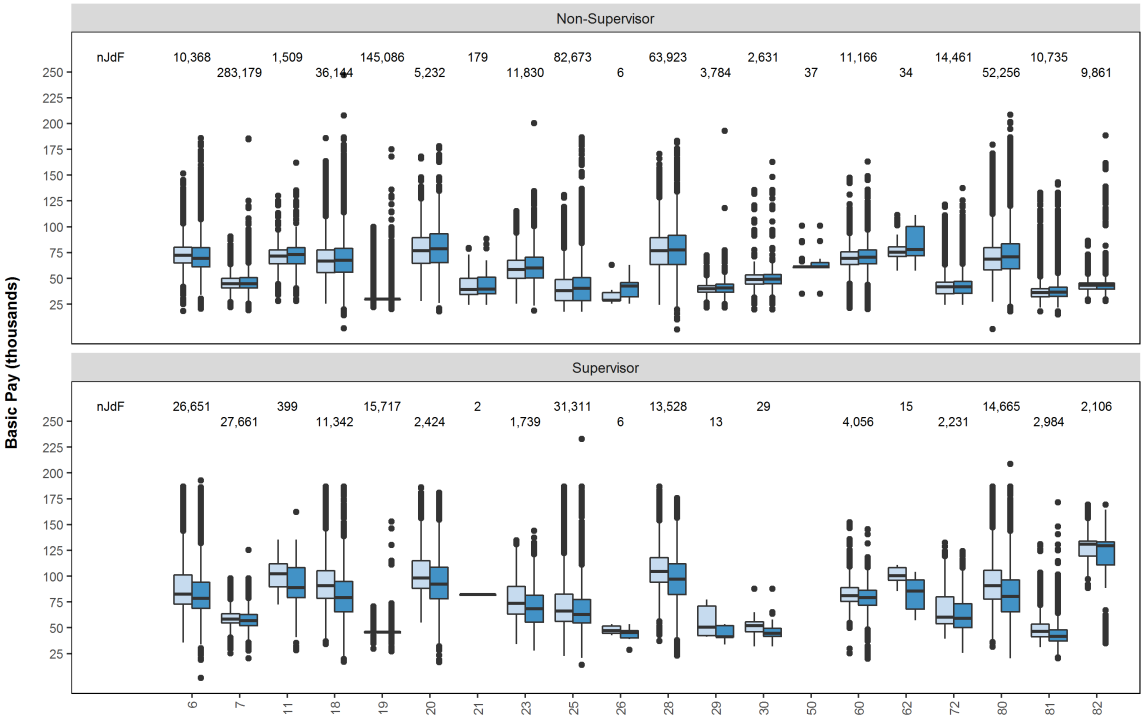


Figure 24: Basic pay distribution by joint professional, supervisory status, college education, and work schedule categories. Department of Veterans Affairs (VA). Dashed line for synthetic data, solid line for authentic.

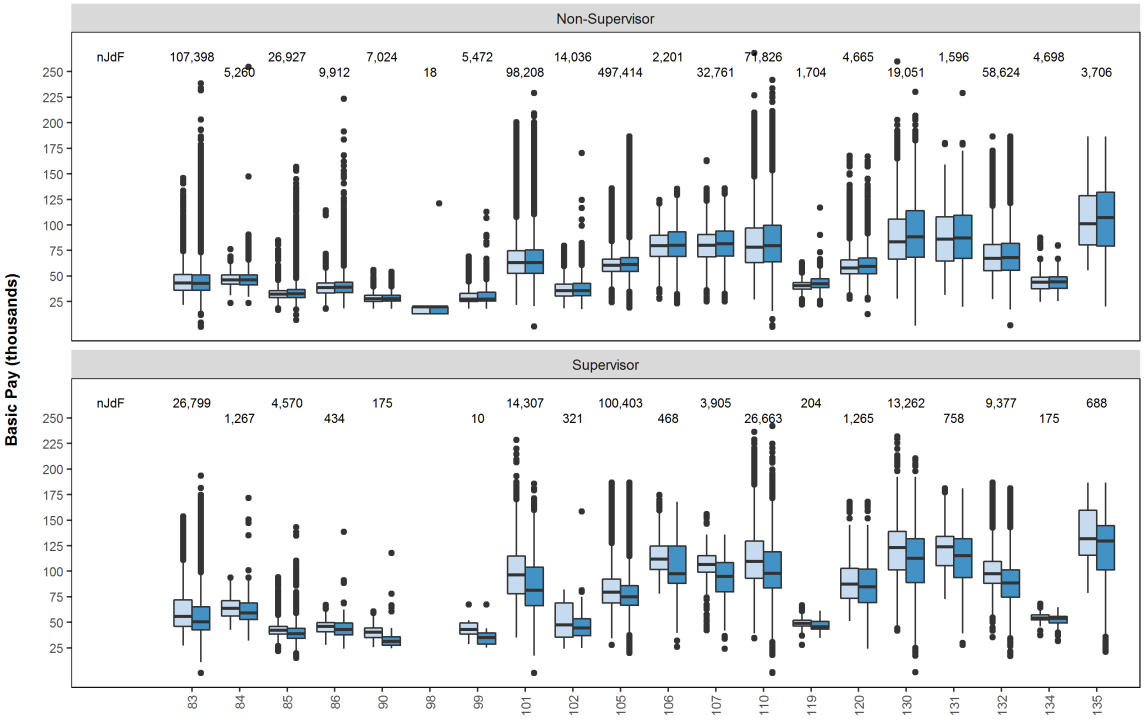
5.3 Distribution of Basic Pay by Occupation and Supervisory Status

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 25 through 27 show box plots for the first 120 occupations in order of occupation code. Trade, or blue collar, occupations begin at code 2500. Figures 28 and 29 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, some difference in pay extremes may be expected.

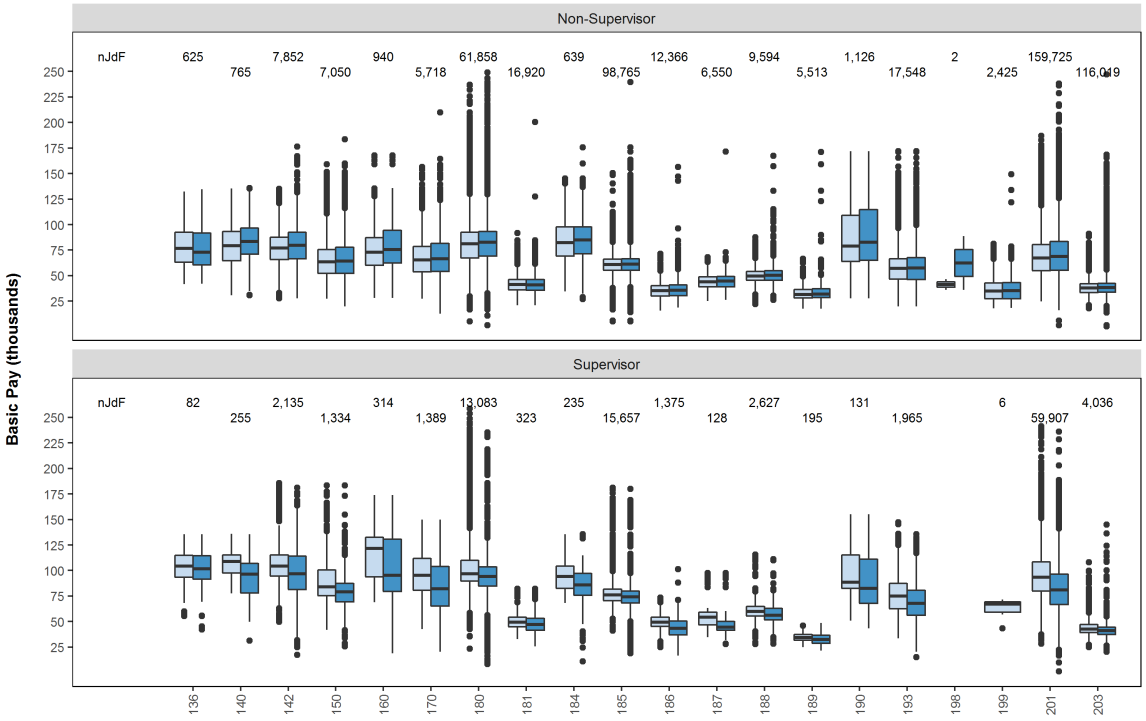


(a) Occupations 0006 through 0082

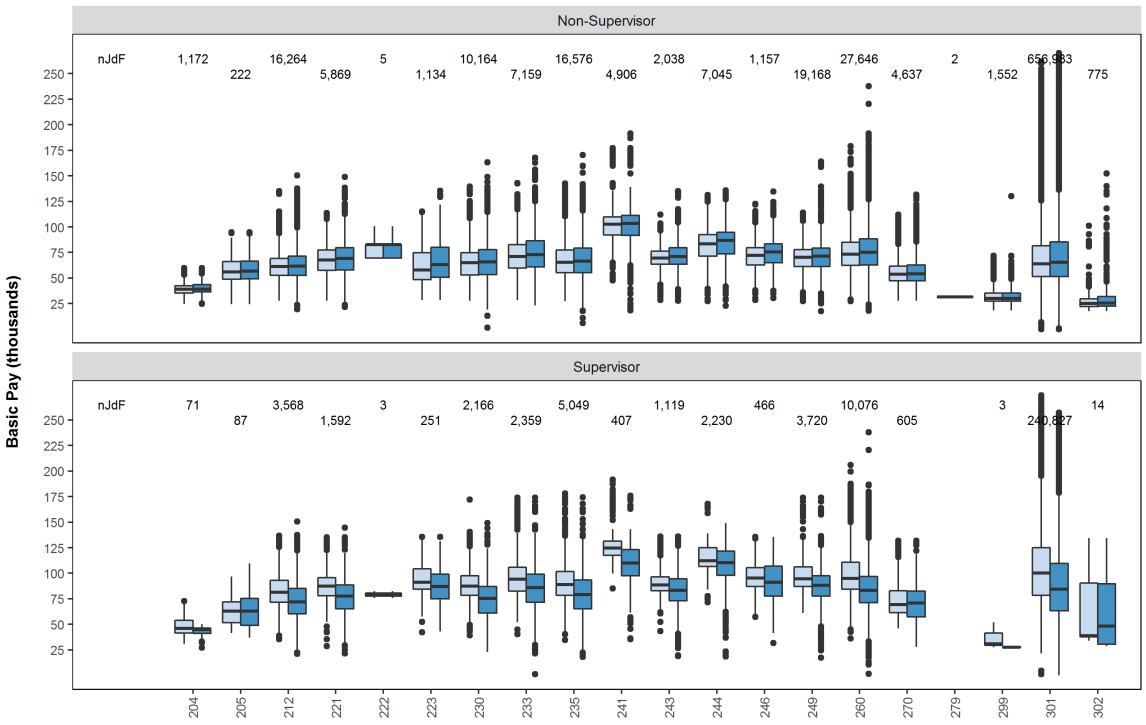


(b) Occupations 0083 through 0135

Figure 25: Basic pay distribution by occupation and supervisor status. All agencies combined. Authentic boxes on left, synthetic on right.

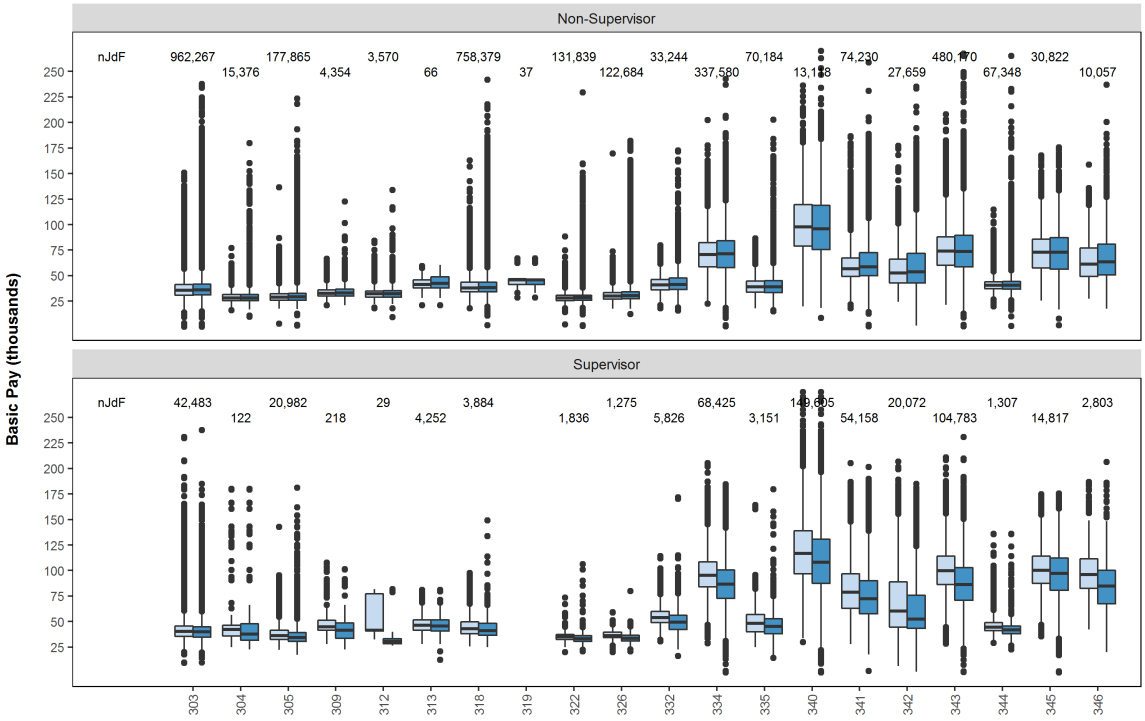


(a) Occupations 0136 through 0203

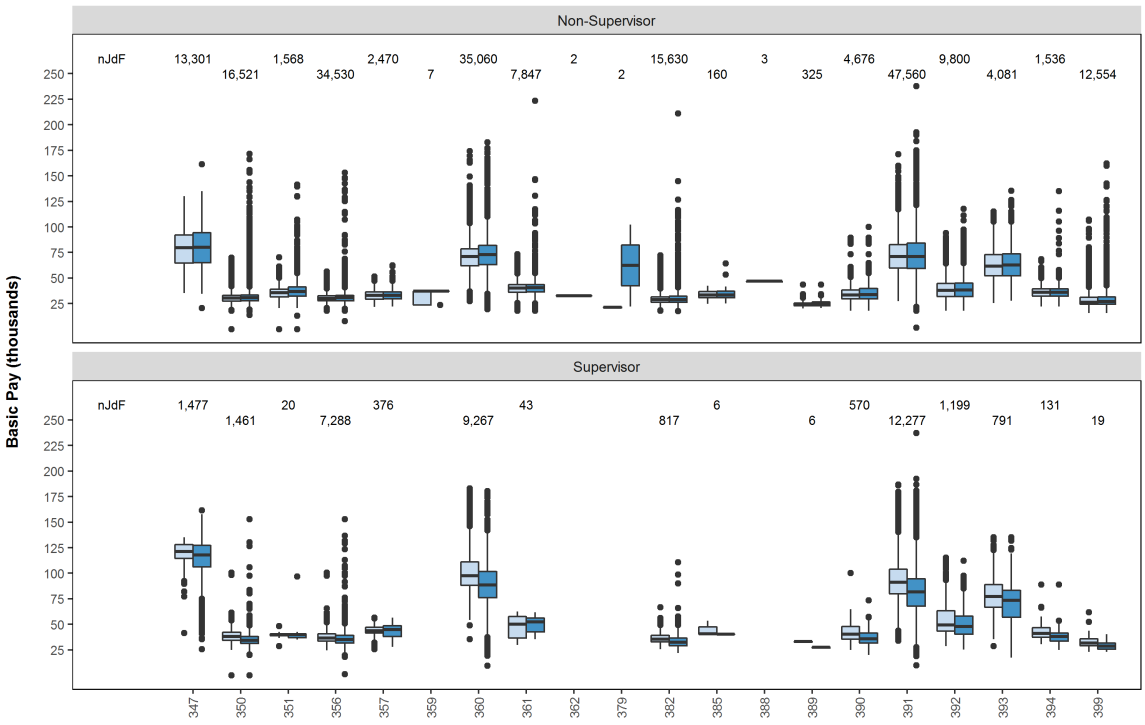


(b) Occupations 0204 through 0302

Figure 26: Basic pay distribution by occupation and supervisor status. All agencies combined. Authentic boxes on left, synthetic on right.

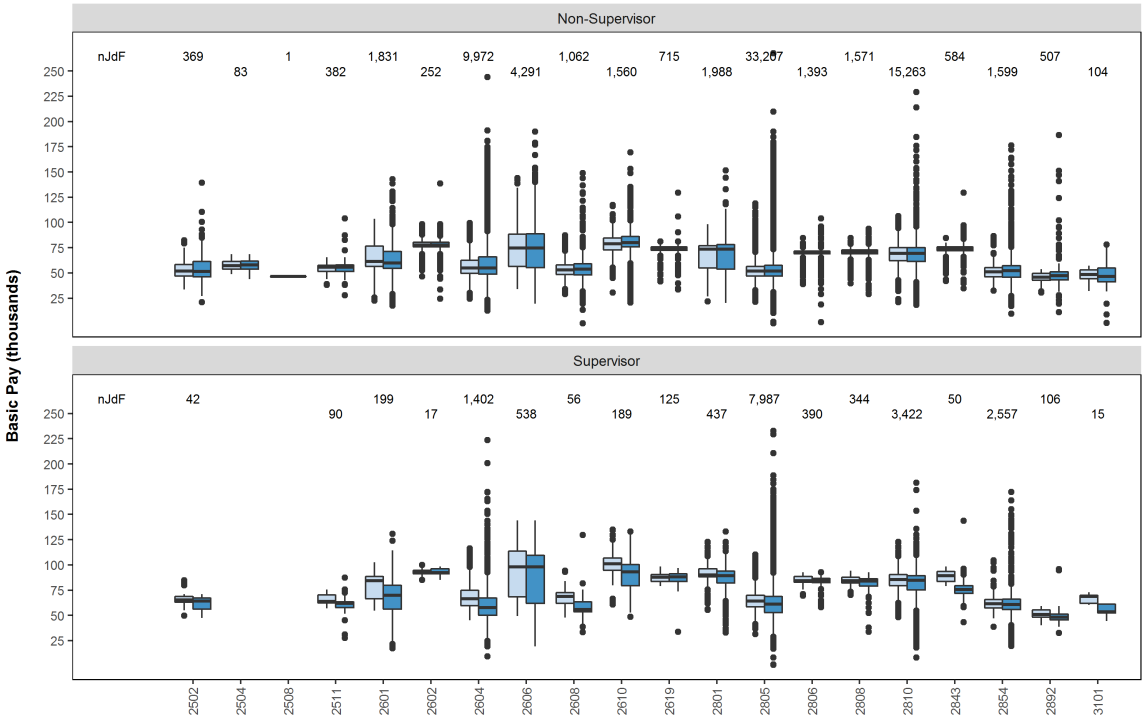


(a) Occupations 0303 through 0346

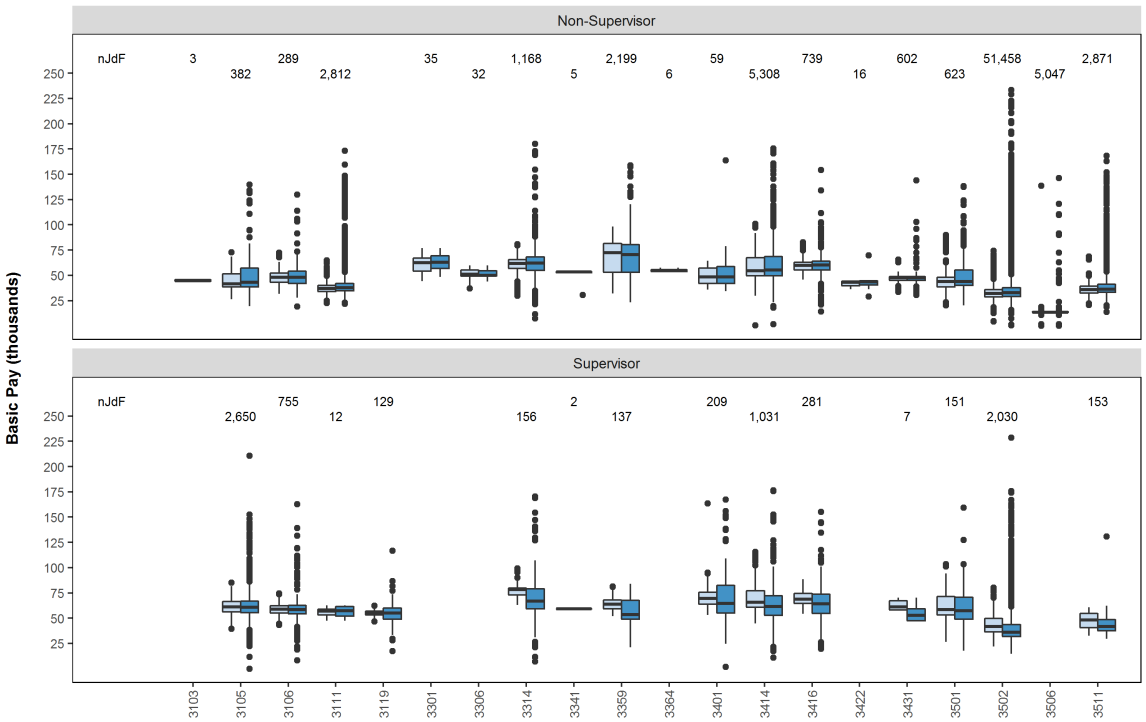


(b) Occupations 0347 through 0399

Figure 27: Basic pay distribution by occupation and supervisor status. All agencies combined. Authentic boxes on left, synthetic on right.

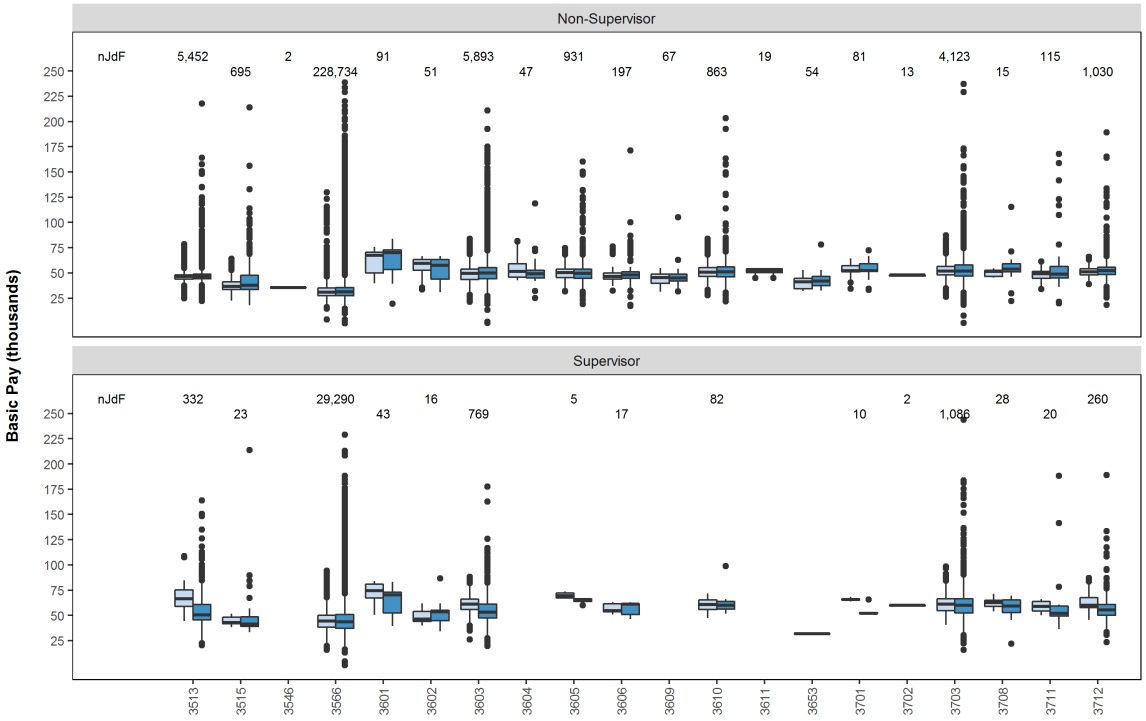


(a) Occupations 2502 through 3101 (trades)

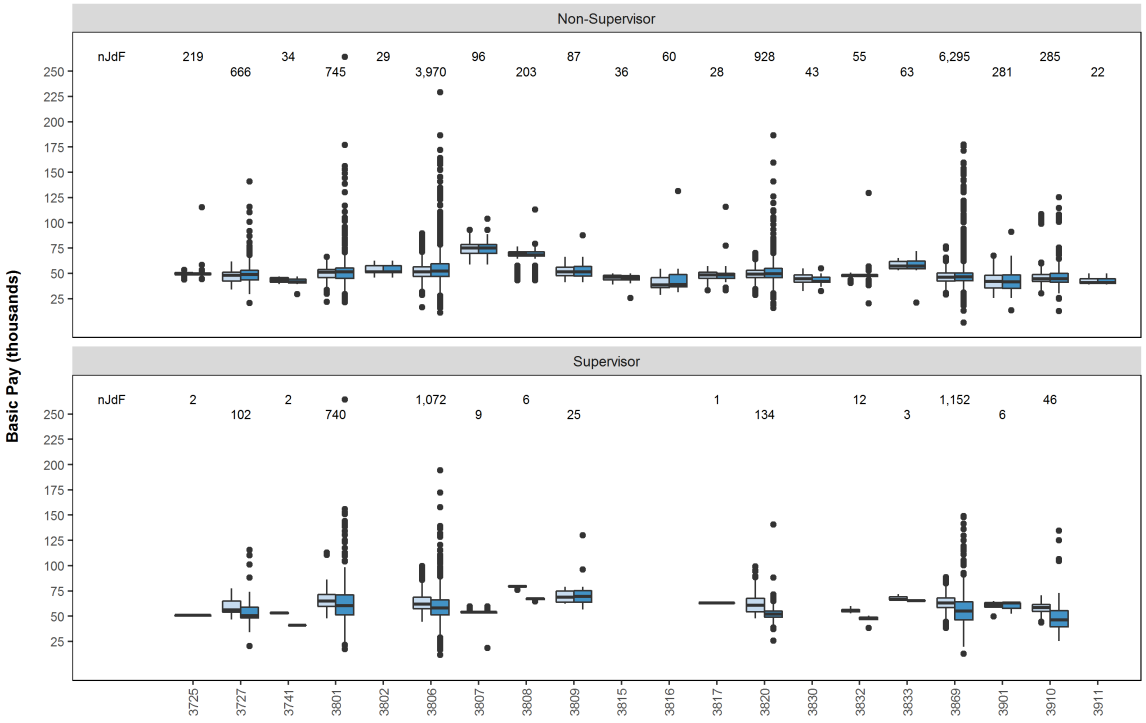


(b) Occupations 3103 through 3511 (trades)

Figure 28: 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 29: Basic pay distribution by occupation and supervisor status. All agencies combined. Authentic boxes on left, synthetic on right.

5.4 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 30 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.

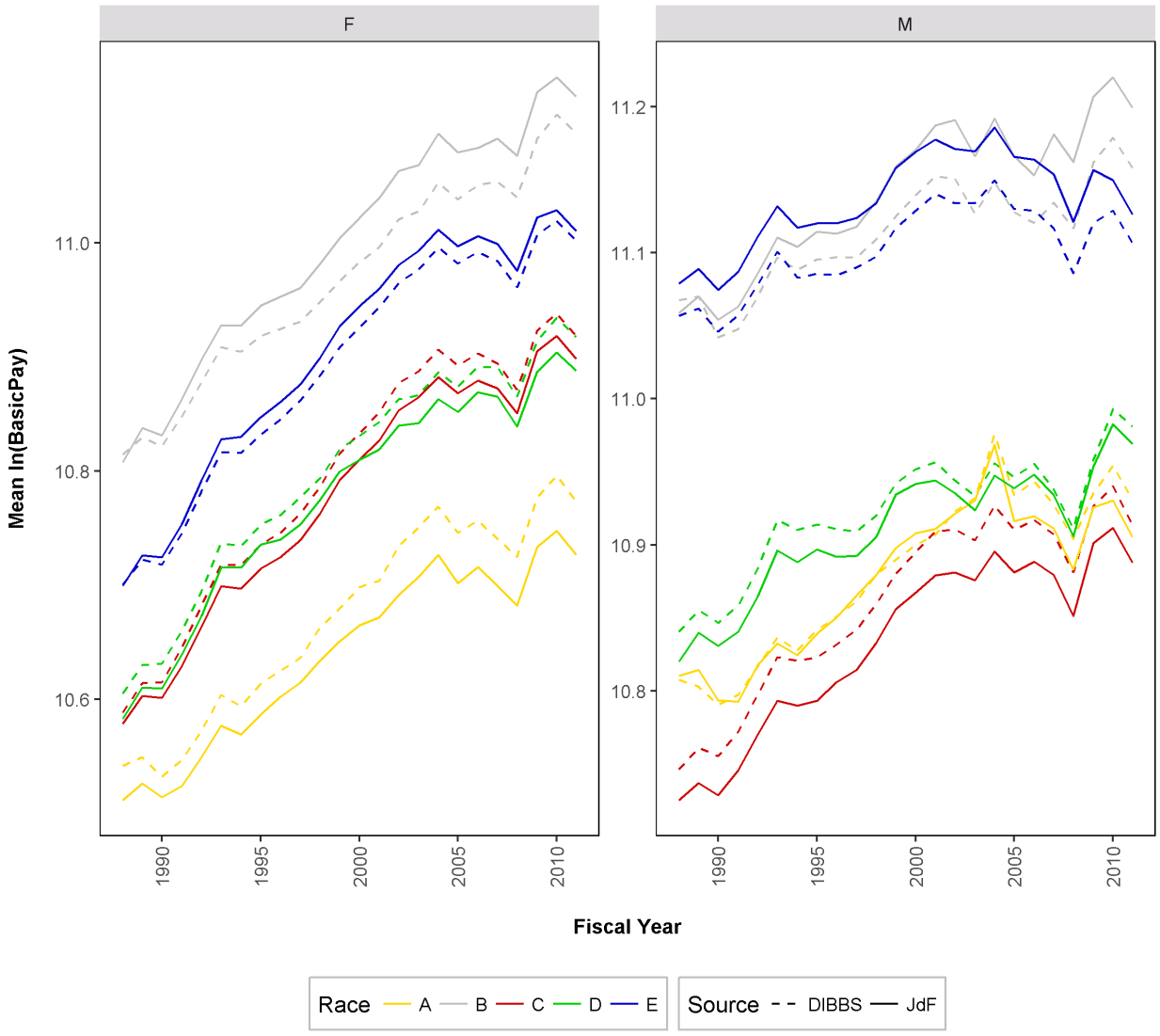


Figure 30: Mean basic pay by sex, race, and year. Female on left, male on right. Race codes: A = Native American, B = Asian, C = black, D = Hispanic, E = white. Dashed line for synthetic data, solid line for authentic.

6 Distribution of Gender

Disparities in pay, job placement, and promotion with respect to gender are important and common topics in human capital research. For synthetic data to produce meaningful results, proportion observations by gender must reflect those observed in corresponding authentic data. This section compares proportions by gender using joint combinations of important organizational and human capital variables.

6.1 Gender Proportion by Race, Education, and Year

Accurate proportion observations by gender is critical in reproducing authentic results with models that include gender as an independent variable. Figure 31 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.

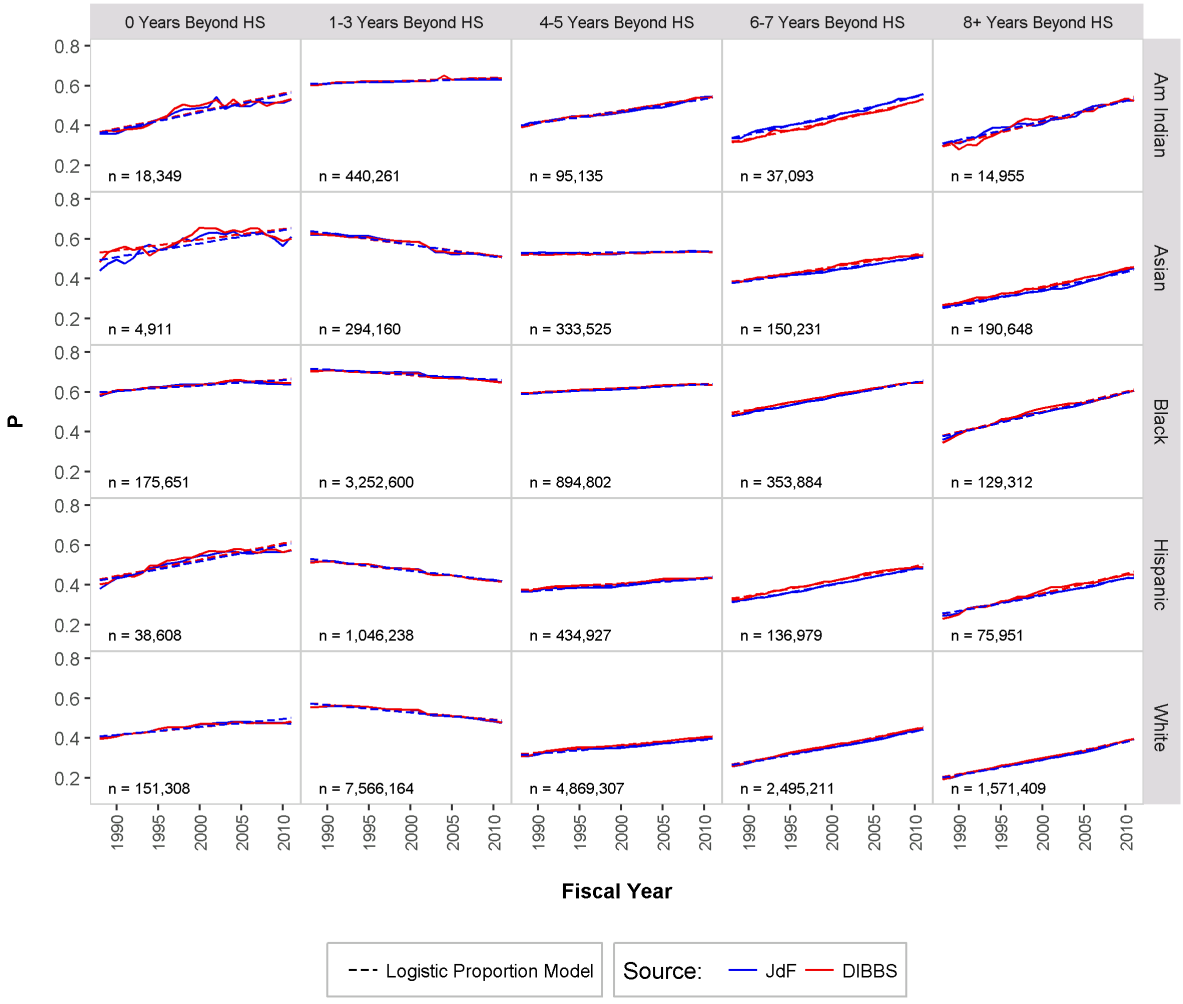


Figure 31: Proportion female observations by race, education, and year. Fitted lines are logistic regression estimates.

6.2 Gender Proportion by Race, Age, and Year

Figures 32 through 36, 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.

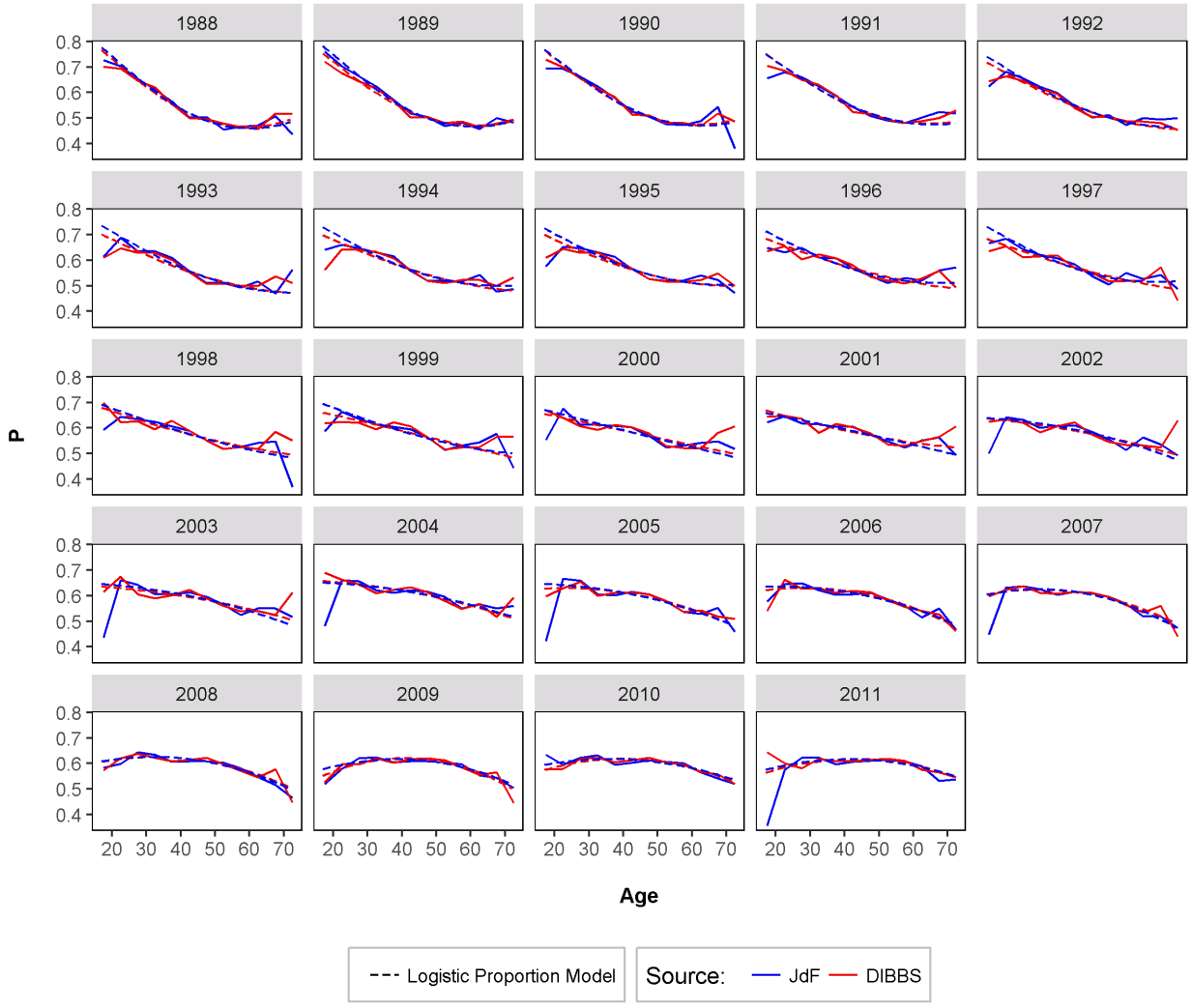


Figure 32: Proportion female observations by education and year. Race Native American. Fitted lines are logistic regression estimates.

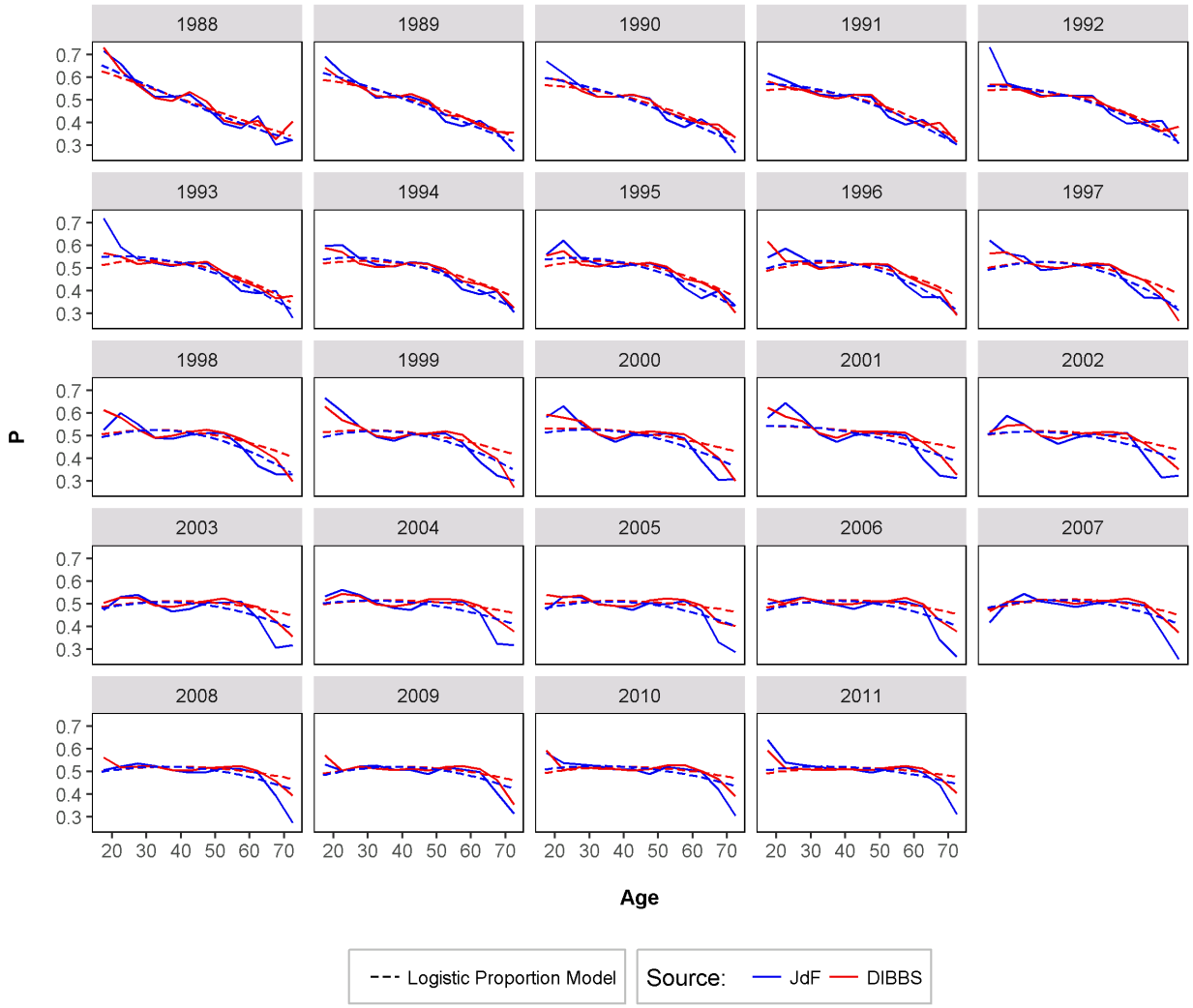


Figure 33: Proportion female observations by education and year. Race Asian. Fitted lines are logistic regression estimates.

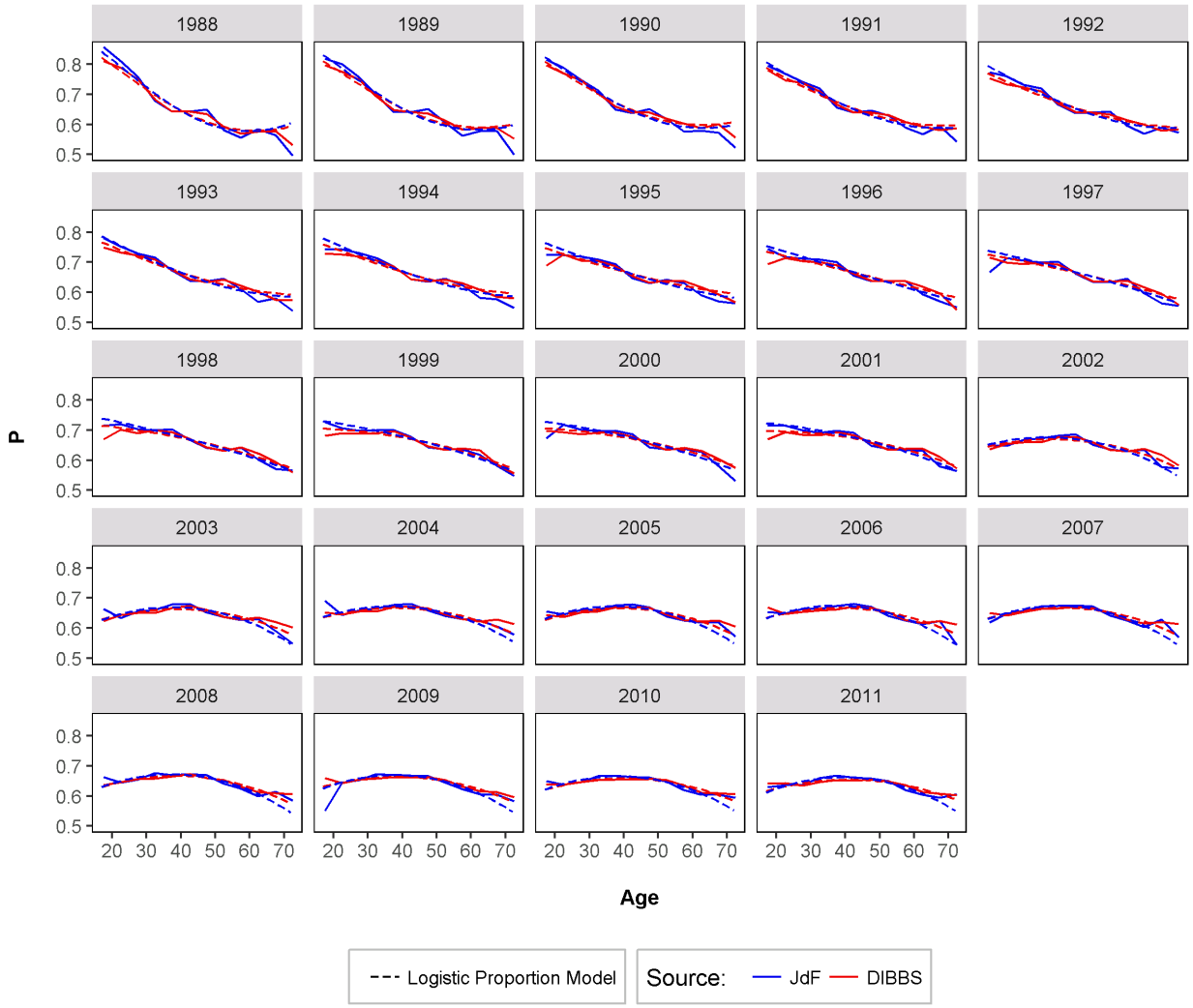


Figure 34: Proportion female observations by education and year. Race black. Fitted lines are logistic regression estimates.

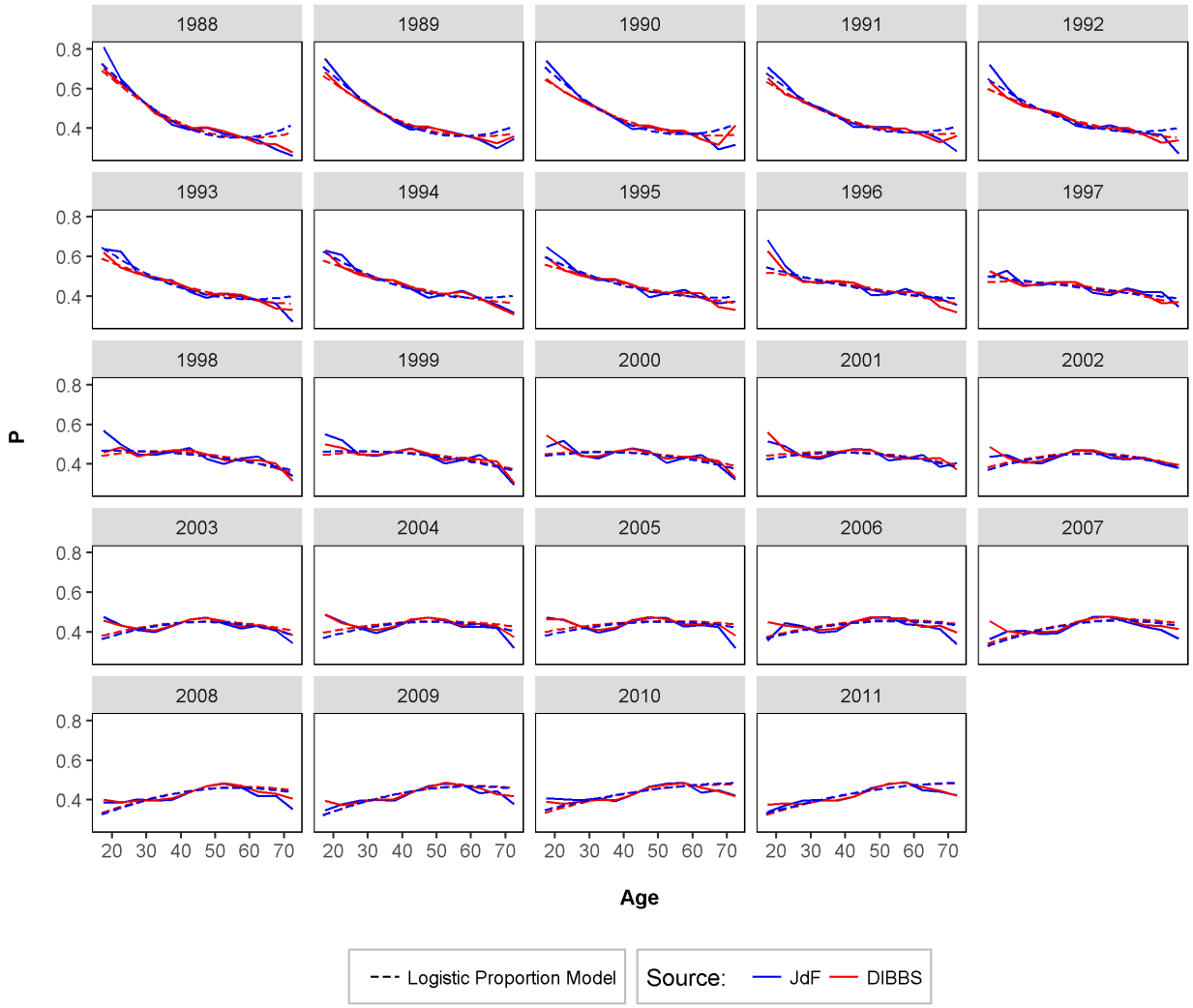


Figure 35: Proportion female observations by education and year. Race Hispanic. Fitted lines are logistic regression estimates.

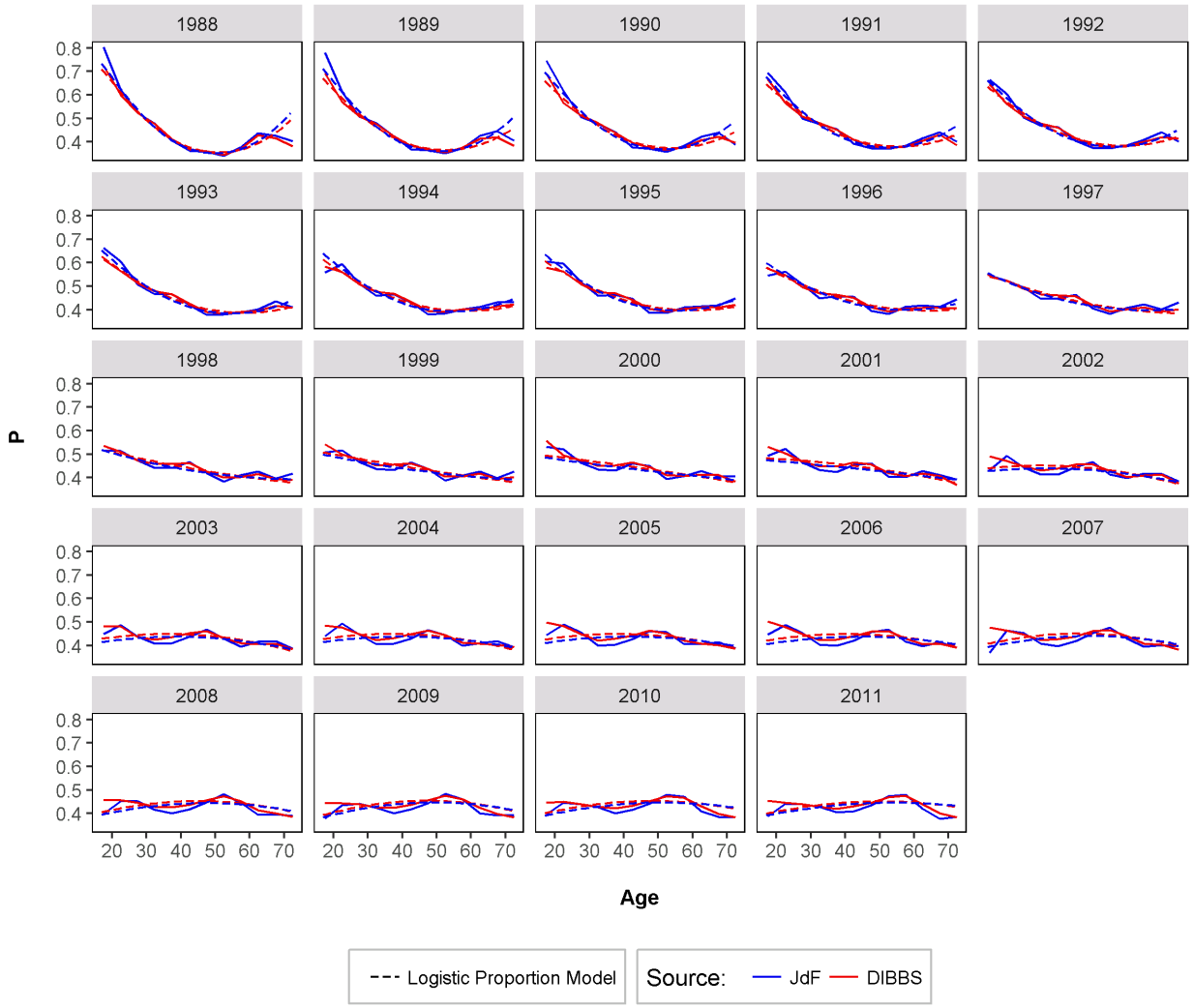
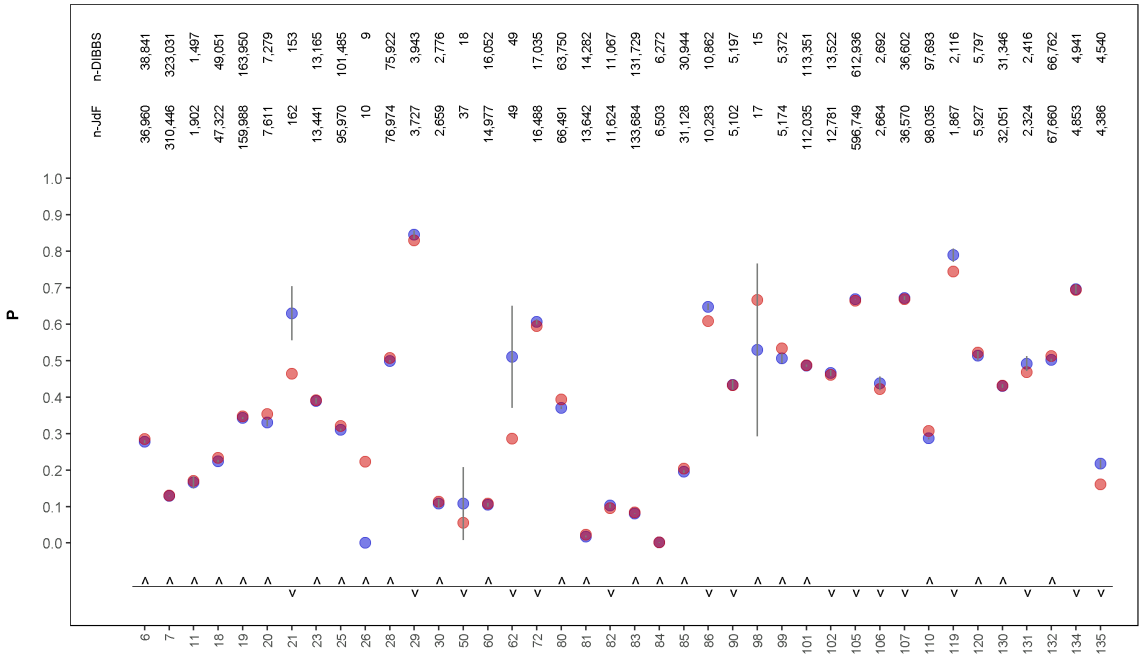


Figure 36: Proportion female observations by education and year. Race white. Fitted lines are logistic regression estimates.

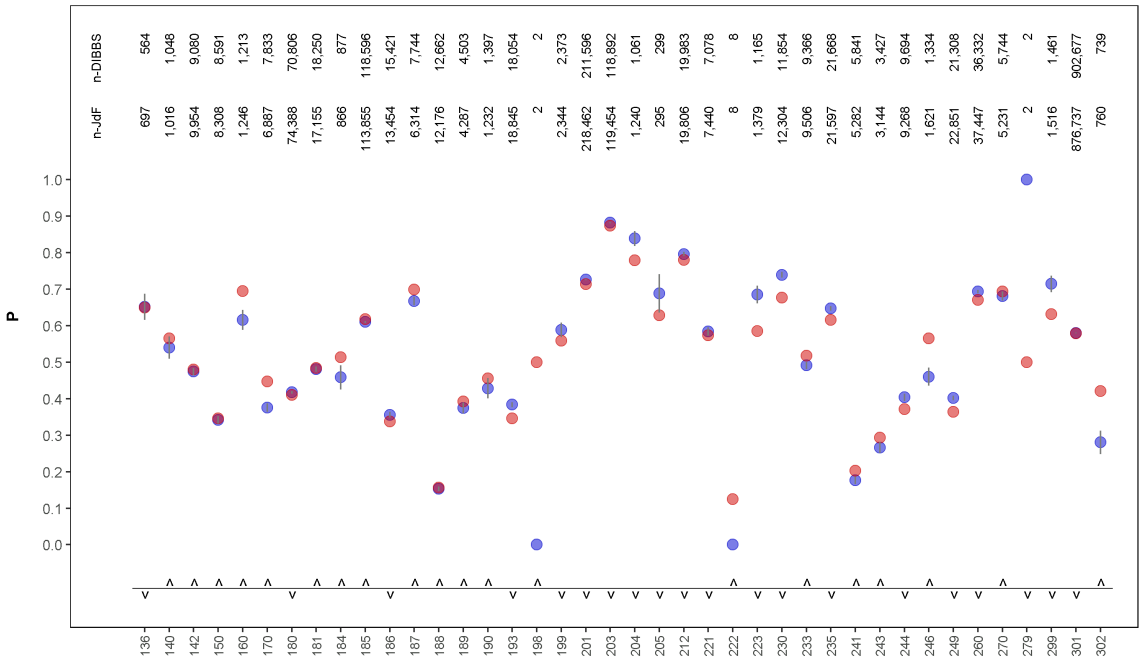
6.3 Gender Proportion by Occupation

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 37 and 38 compare proportion female for the first 120 occupation codes. Figures 39 and 40 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.

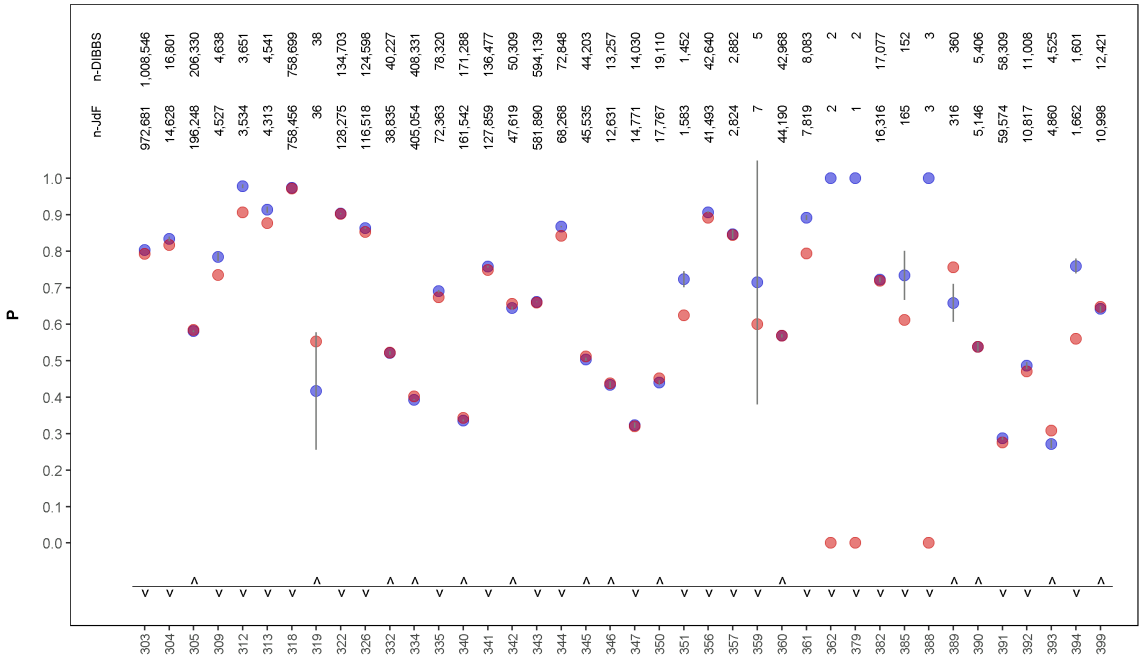


(a) Occupations 0006 through 0135

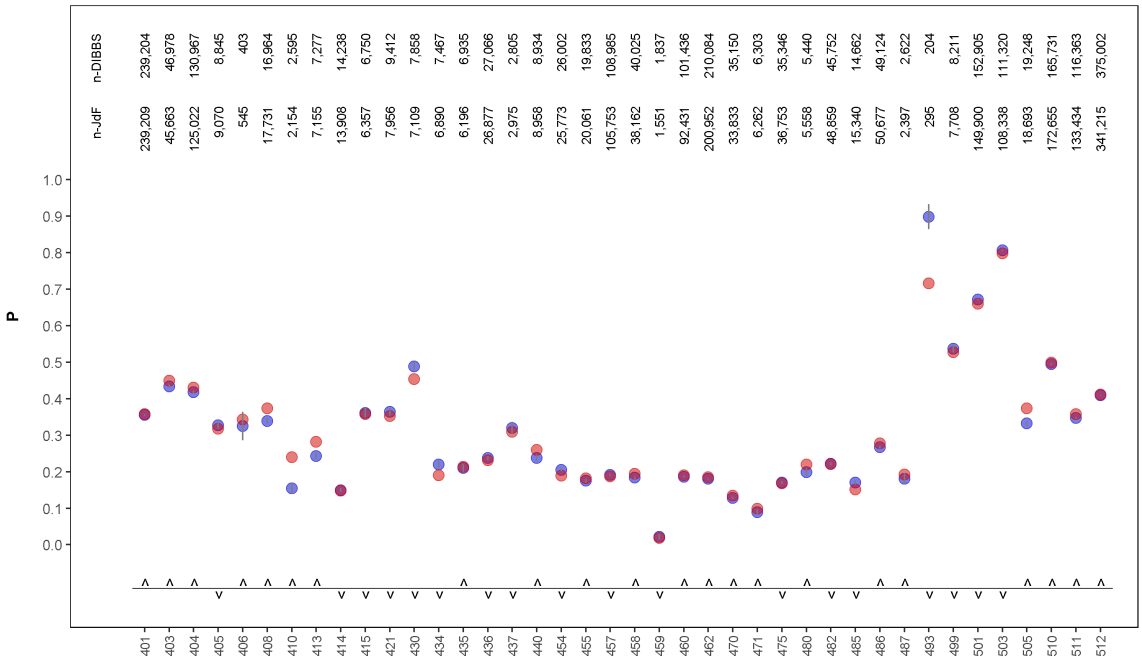


(b) Occupations 0136 through 0302

Figure 37: Proportion female observations by occupation. All agencies combined. One synthetic and one authentic point per occupation.

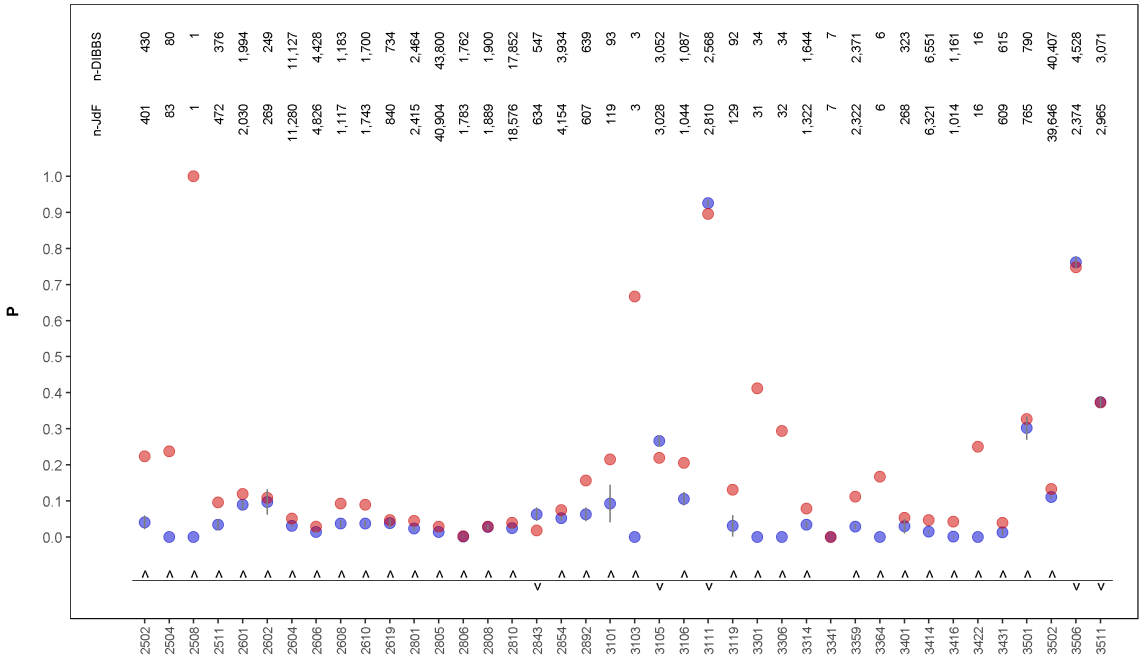


(a) Occupations 0303 through 0399

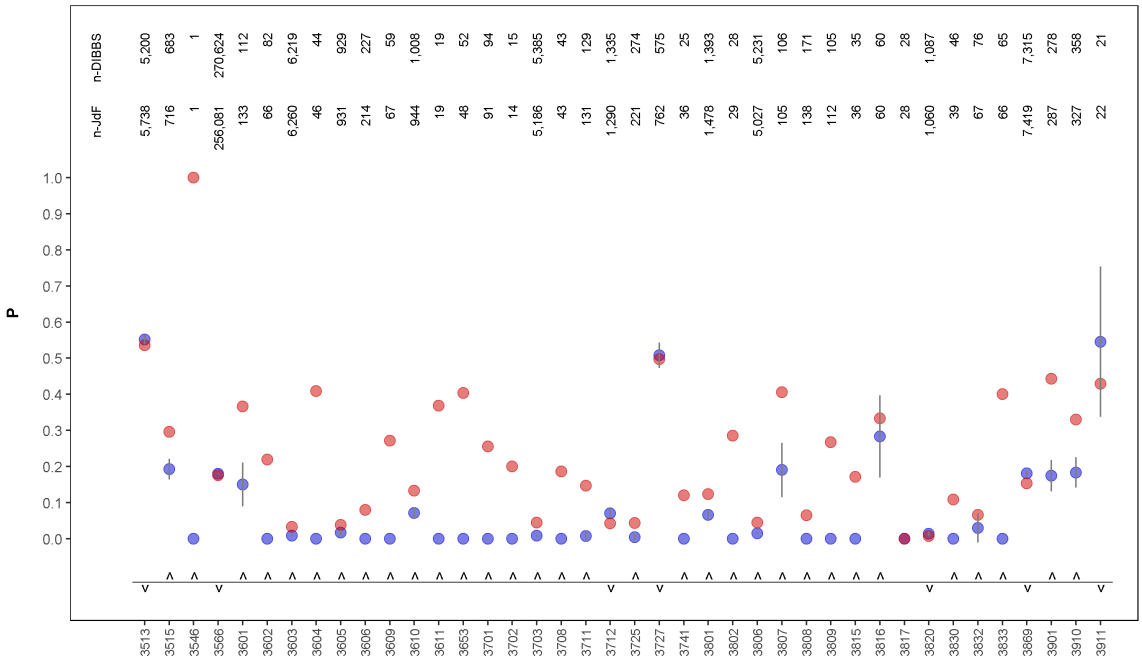


(b) Occupations 0401 through 0512

Figure 38: Proportion female observations by occupation. All agencies combined. One synthetic and one authentic point per occupation.

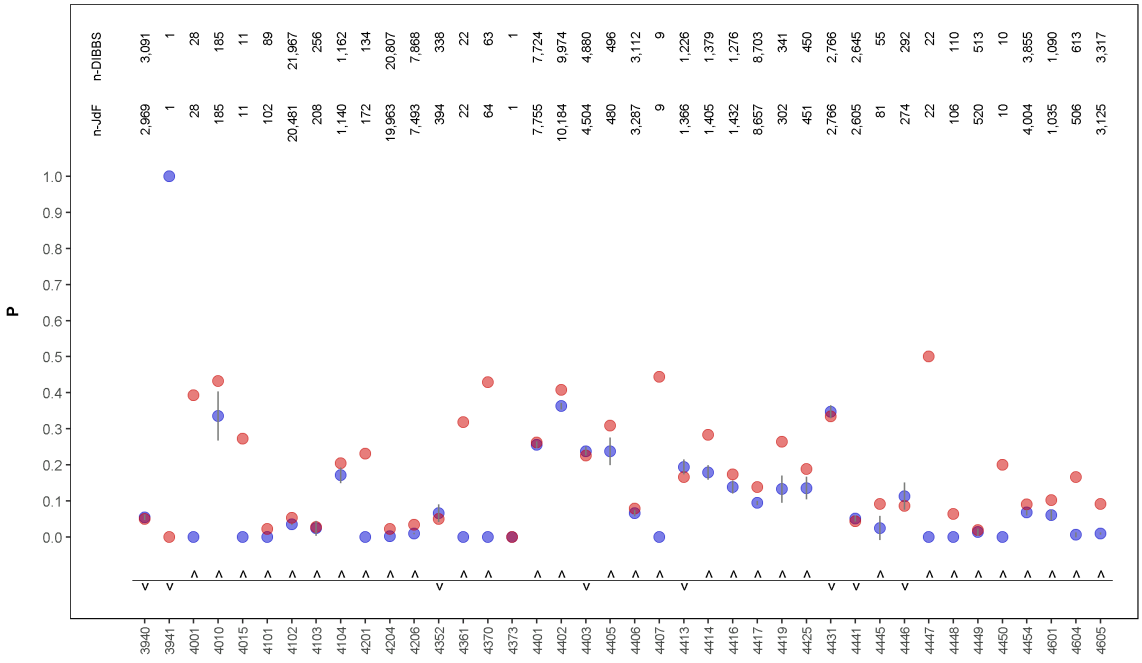


(a) Occupations 2502 through 3511 (trades)

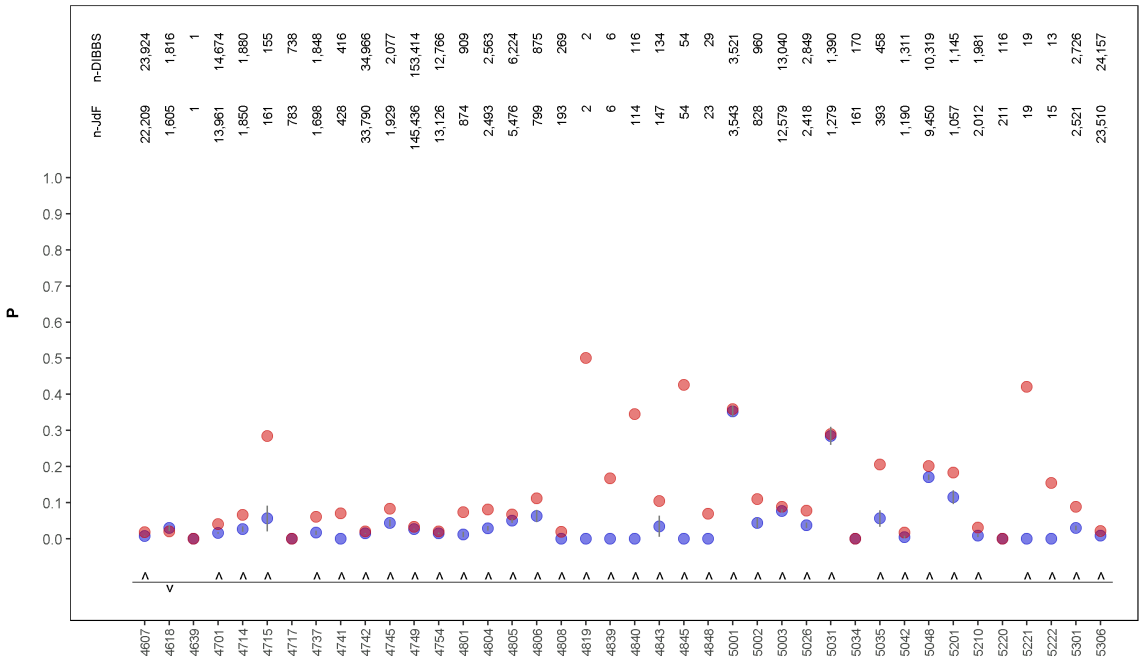


(b) Occupations 3513 through 3911 (trades)

Figure 39: 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 40: Proportion female observations by occupation. All agencies combined. One synthetic and one authentic point per occupation.

6.4 Occupation Gender Proportion Kernel Density

Figure 41 superimposes synthetic and authentic kernel density plots of proportion female employees by occupation.

Observations: There is some discrepancy in density for synthetic proportions near 0 and above 0.75, which is compensated for near more central proportions. Since very low or very high proportion observations within important identifiers (sex in this case) can promote individual identification, discrepancies in extremes may be expected.

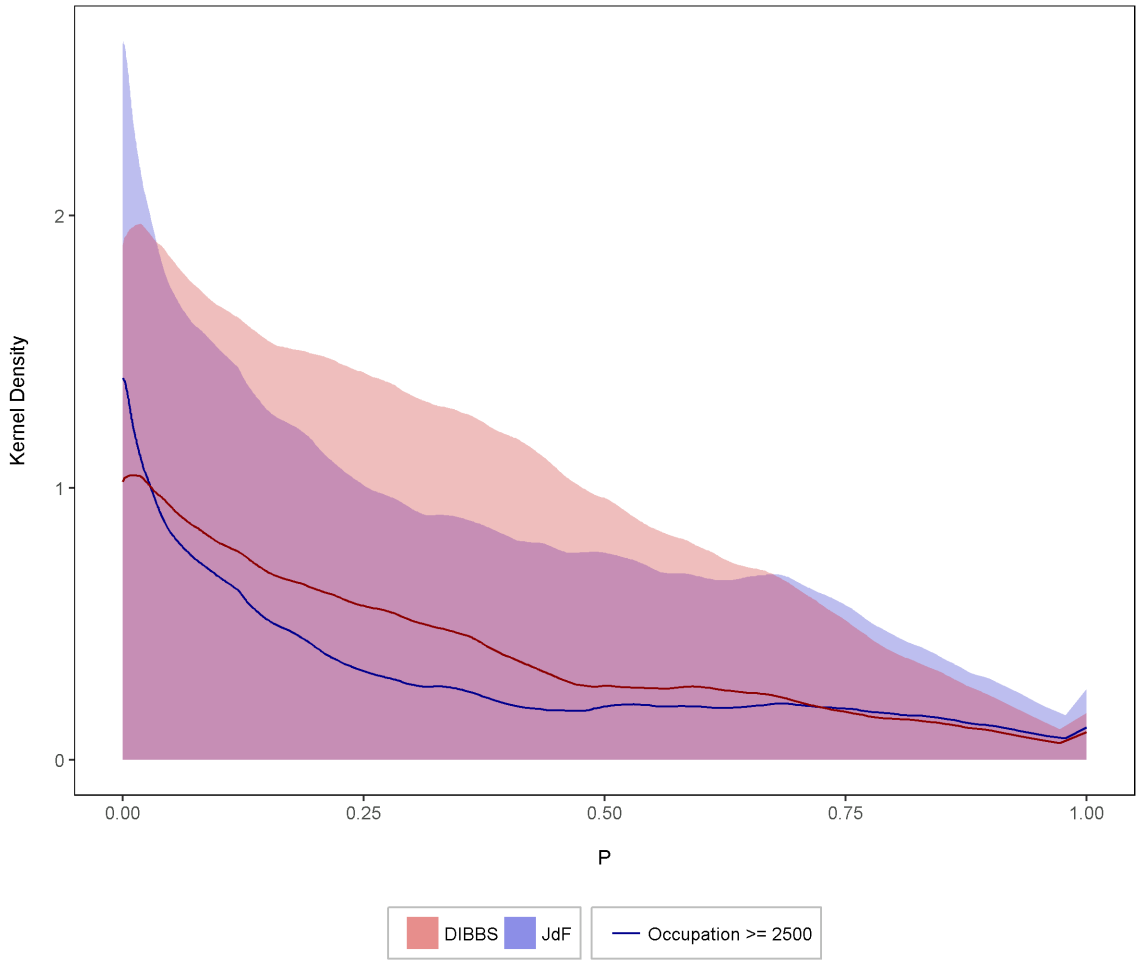


Figure 41: Occupation proportion female kernel density. Synthetic and authentic distributions superimposed. Trade occupations slightly over-represented near zero and above 0.75 in authentic data. Slight discrepancy in synthetic proportions at extremes. Compensated for near central proportions.

6.5 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 42 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.

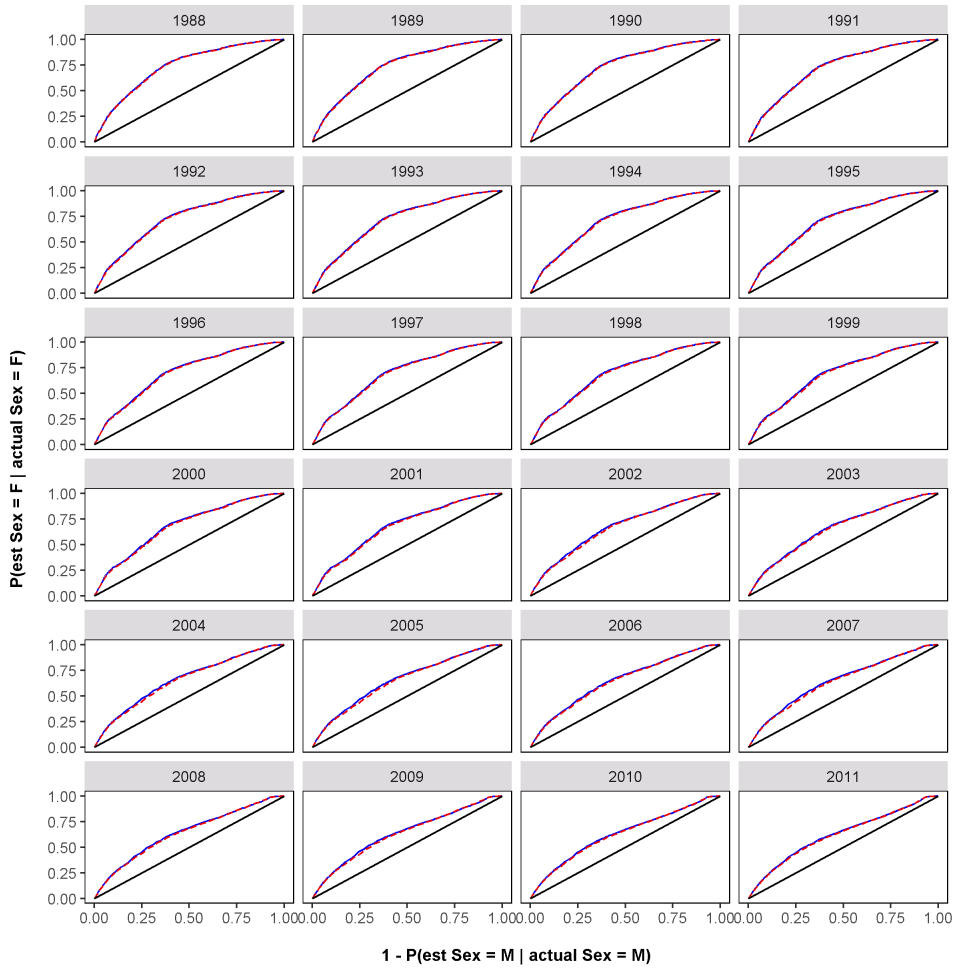


Figure 42: Proportion female race, age, education, occupation classifier ROC curves. One curve per data set per by fiscal. Agreement in classifier accuracy indicated by overlapping curves. Pattern of decreased accuracy as years progress captured in both data sets.

7 Gender Pay Disparity Fixed Effects Models

7.1 Fixed Effects Ordinary Least Squares Regression Model

In their study of pay disparity in the federal government, Alexander Bolton and John de Figueiredo report an effect of gender on difference in basic pay using fixed effects regression models that control for important human capital factors (Bolton and de Figueiredo, 2017). An example model is

$$y = \beta_0 + \beta_f female + \beta_r race + \beta_{ag} age + \beta_{ag^2} age^2 + \beta_{ed} education + \beta_{bur} bureau + \beta_{occ} occupation + \beta_{yr} year \quad (1)$$

where y is the expected value of log(basic pay), given specific levels of race, age, years of education, bureau (proxy for agency), occupation, and year. Using *male* as the sex reference level, $\hat{\boldsymbol{\beta}}_f$, estimated from model (1), measures the difference in pay between females and males after correcting for the remaining variables. Table 2 lists parameter estimates for model (1) fit to full-time, pay plan GS synthetic and authentic observations. Common reference levels were used for both sets (sex male, race white, highest frequency bureau and occupation, and fiscal year 1988). Corresponding standard errors appear in parentheses beneath each estimate.[6] Figure 43 plots synthetic data estimates for model (1) against corresponding authentic data estimates, one point for each estimate. The "X" group includes estimates for sex, race, education, and age. Figures 44, 45, and 46 plot parameter estimates obtained by fitting model (1) to annual subsets of synthetic and authentic data. These are used to identify potential patterns of change in parameter estimate association throughout the study period.

Observations (table): Although some discrepancies exist between estimates from the two data sets, the close proximity of $\hat{\boldsymbol{\beta}}_f$ estimates confirms the utility of synthetic data results for gender pay disparity research.[7] In their analysis, Bolton and de Figueiredo conclude that the reported pay disparity of approximately -0.03 is significant. Using the synthetic data, a researcher would measure very similar disparity and, presumably, reach a similar conclusion of significance.

Observations (figure 43): Points generally lie near the reference line of slope 1.0, indicating overall agreement of synthetic and authentic parameter estimates. Light dots indicate low frequency bureaus and occupations and represent most of the large deviations from reference lines.

Observations (annual figures): Points generally lie near reference lines in all years for the three groups. Low frequency bureaus and occupations (light dots) account for deviations from reference. There does not appear to be any pattern of change in association of estimates between data sets throughout the period.

[6] Although Bolton and de Figueiredo compute robust standard errors clustered about employee, for simplicity we ignore potential heteroskedasticity of residuals and present here (homoskedastic residual based) standard errors estimated using the diagonal of the inverse of the covariance matrix.

[7] Some deviations in parameter estimates may appear large with respect to corresponding standard errors. However, due to relatively large observation counts, n, standard errors are small, implying that the data sets are more like populations than samples, which magnifies deviations and calls into question customary notions of parameter estimate distribution.

Table 2: Parameter estimates from gender pay disparity fixed effects model

Parameter Estimate	OPM	DIBBS
$\hat{\boldsymbol{\beta}}_0$	9.3973	9.6167
	(7.19e-04)	(7.23e-04)
$\hat{\boldsymbol{\beta}}_f$	-0.0319	-0.0296
	(1.10e-04)	(1.17e-04)
$\hat{\boldsymbol{\beta}}_{age}$	0.0352	0.0269
	(2.93e-05)	(2.85e-05)
$\hat{\boldsymbol{\beta}}_{age^2}$	-0.0003	-0.0002
	(3.29e-07)	(3.19e-07)
$\hat{\boldsymbol{\beta}}_{ed}$	0.0186	0.0122
	(2.69e-05)	(2.78e-05)
$\hat{\boldsymbol{\beta}}_{raceAmInd}$	-0.0370	-0.0161
	(4.03e-04)	(4.12e-04)
$\hat{\boldsymbol{\beta}}_{raceAsian}$	-0.0320	-0.0253
	(2.53e-04)	(2.70e-04)
$\hat{\boldsymbol{\beta}}_{raceBlack}$	-0.0095	-0.0057
	(1.25e-04)	(1.34e-04)
$\hat{\boldsymbol{\beta}}_{raceHisp}$	-0.0202	-0.0137
	(1.85e-04)	(1.99e-04)
$\hat{\boldsymbol{\beta}}_{bureau}$		332 fixed effect levels
$\hat{\boldsymbol{\beta}}_{occupation}$		466 fixed effect levels
$\hat{\boldsymbol{\beta}}_{year}$		23 fixed effect levels
n	18,165,075	18,714,815

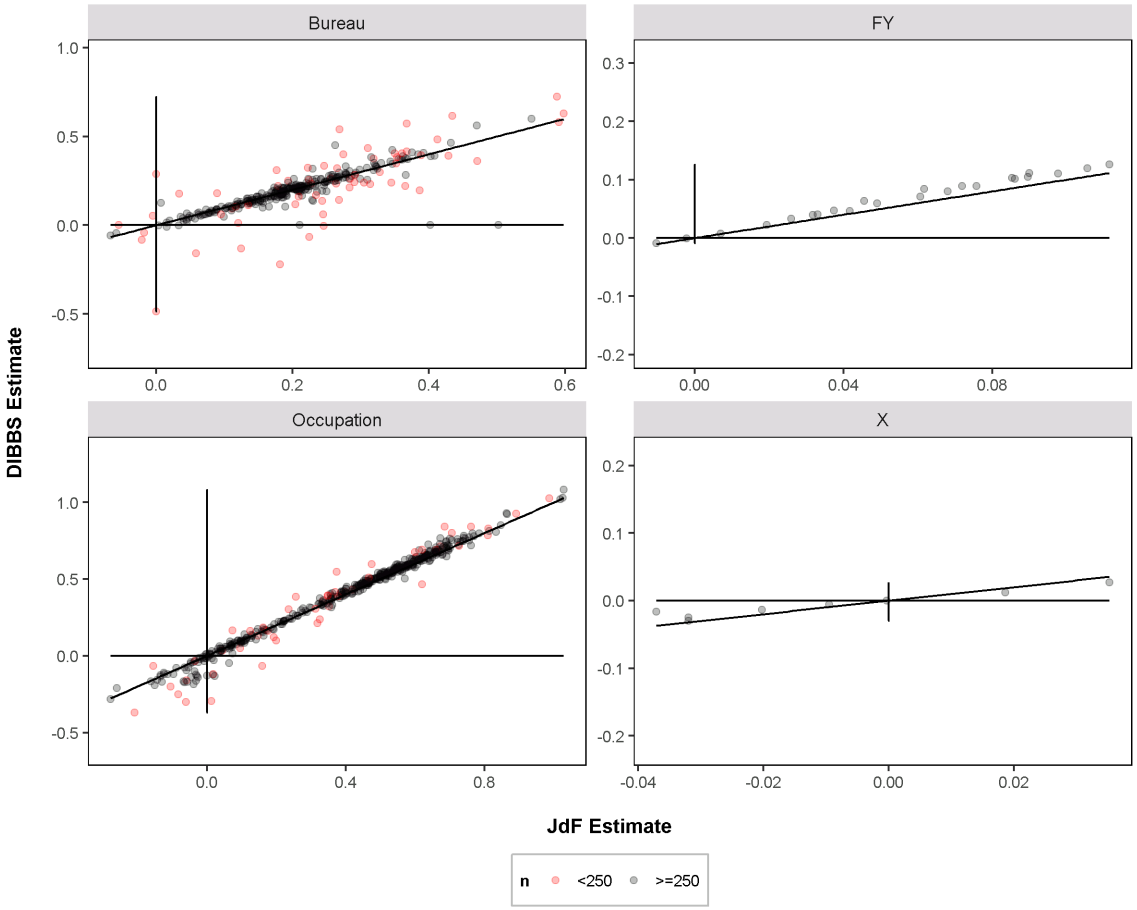


Figure 43: Pay disparity fixed effects regression model parameter estimates. X group includes estimates for sex, race, education, and age. Points lie near reference line of slope 1.0, indicating general agreement of synthetic and authentic parameter estimates. Light dots represent low frequency bureaus and occupations. They are associated with large deviations from reference line.

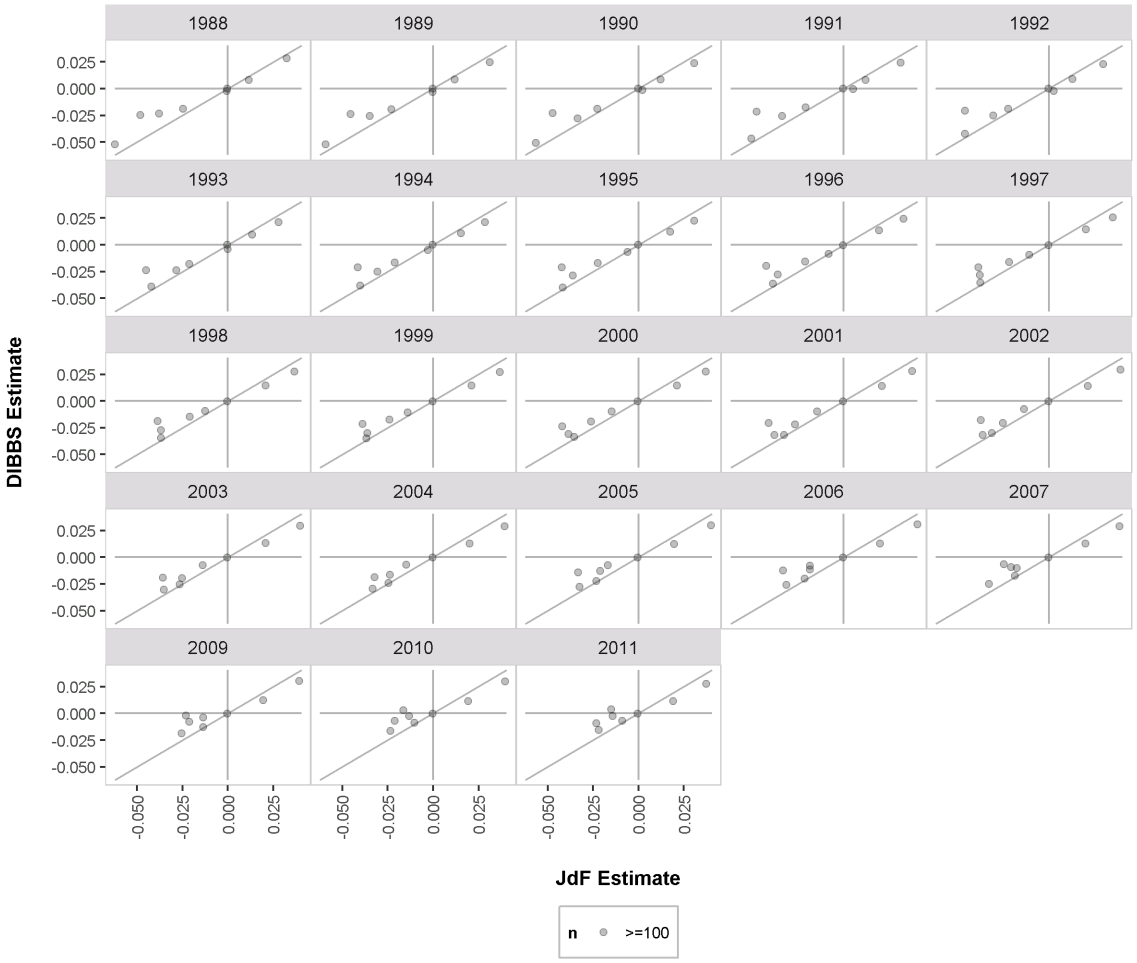


Figure 44: Sex (female), race, age, age^2, and years of education parameter estimates from annual pay disparity fixed effects regression model. Points near reference line, indicating similarity. No pattern of change throughout study period.

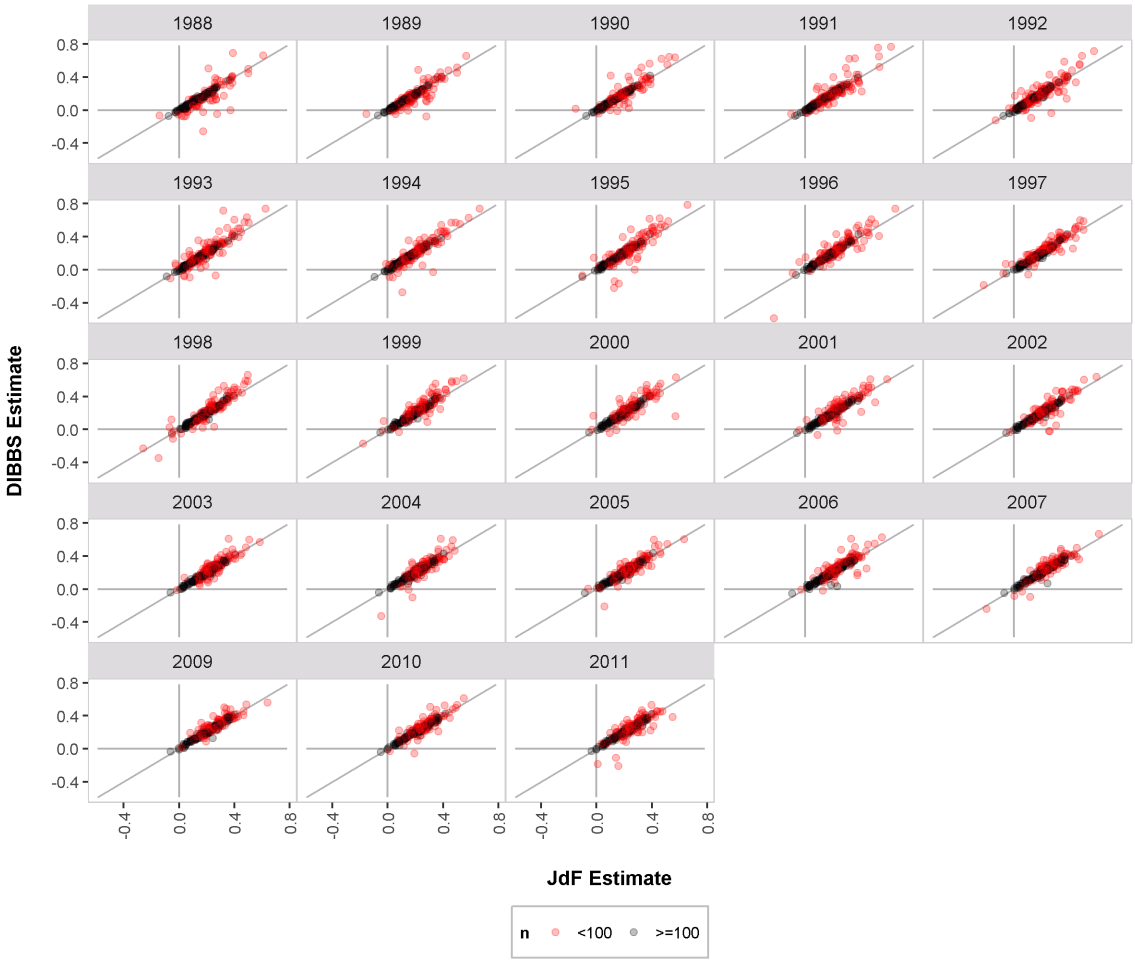


Figure 45: Bureau (agency) parameter estimates from annual pay disparity fixed effects regression model. Points near reference line, indicating similarity. Light points away from reference represent low frequency bureaus. No pattern of change throughout study period.

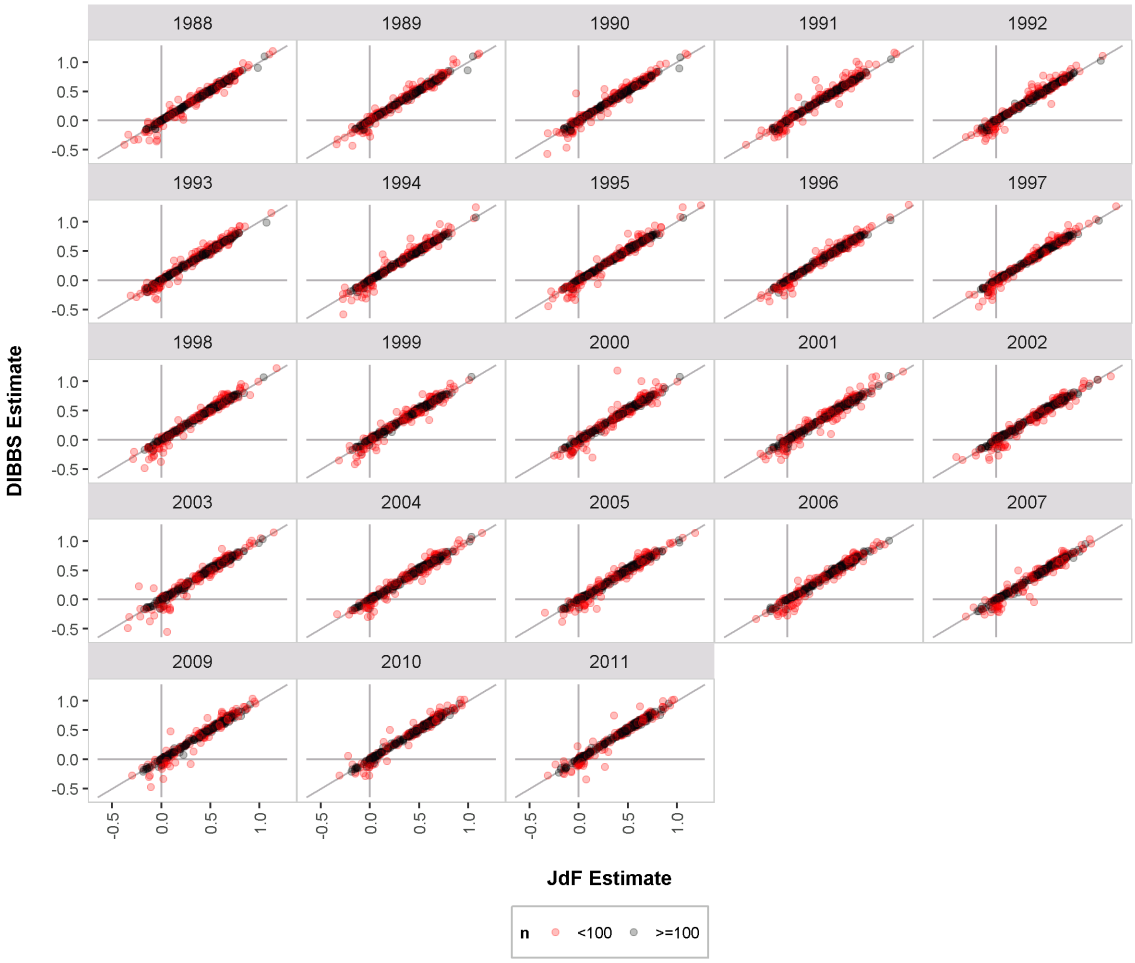


Figure 46: Occupation parameter estimates from annual pay disparity fixed effects regression model. Points near reference line, indicating similarity. Light points away from reference represent low frequency bureaus. No pattern of change throughout study period.

7.2 Fixed Effects Quantile Regression Model

Disparity in pay attributed to gender is an important and common topic in human capital research. Ordinary least squares regression estimates the effect of gender on expected values of difference in pay, but also of interest is the effect of gender on estimates of particular quantiles, pay values below which a given proportion of observations reside. Of additional interest is the change in this effect over time, if any. An example model used to estimate the effect of gender on pay quantiles for a given year, controlling for race, age, education, agency, and occupation is

$$y = \beta_0 + \beta_s sex + \beta_r race + \beta_{ag} age + \beta_{ag^2} age^2 + \beta_{ed} education + \beta_{agcy} agency + \beta_{occ} occupation \qquad (2)$$

where y is a particular quantile of log(basic pay). Figure 47 plots estimates of gender effect ($\hat{\boldsymbol{\beta}}_s$) from model (2) fit to annual subsets of observations for quantiles 0.1, 0.5, and 0.9.

Observations: Estimates of gender effect ($\hat{\boldsymbol{\beta}}$) tend to be greater for quantiles 0.5 and 0.9 than for 0.1 in all years. This indicates a tendency, for median and high pay positions, of greater disparity between women and men with common occupation, education, and experience (age) profiles and, although with some systematic difference to authentic estimates, estimates from the synthetic data identify this important feature. Additionally, the gradual trend toward parity ($\hat{\boldsymbol{\beta}}_s$ estimates approach 0) as years advance that is observed in authentic data is also represented in the synthetic data. Although parity appears to have been achieved in the final year for quantile 0.1, it remains at approximately -0.01 and -0.02 for quantiles 0.5 and 0.9, respectively, indicating an increasing lag of disparity as pay level increases. This important result is also apparent in the synthetic data.

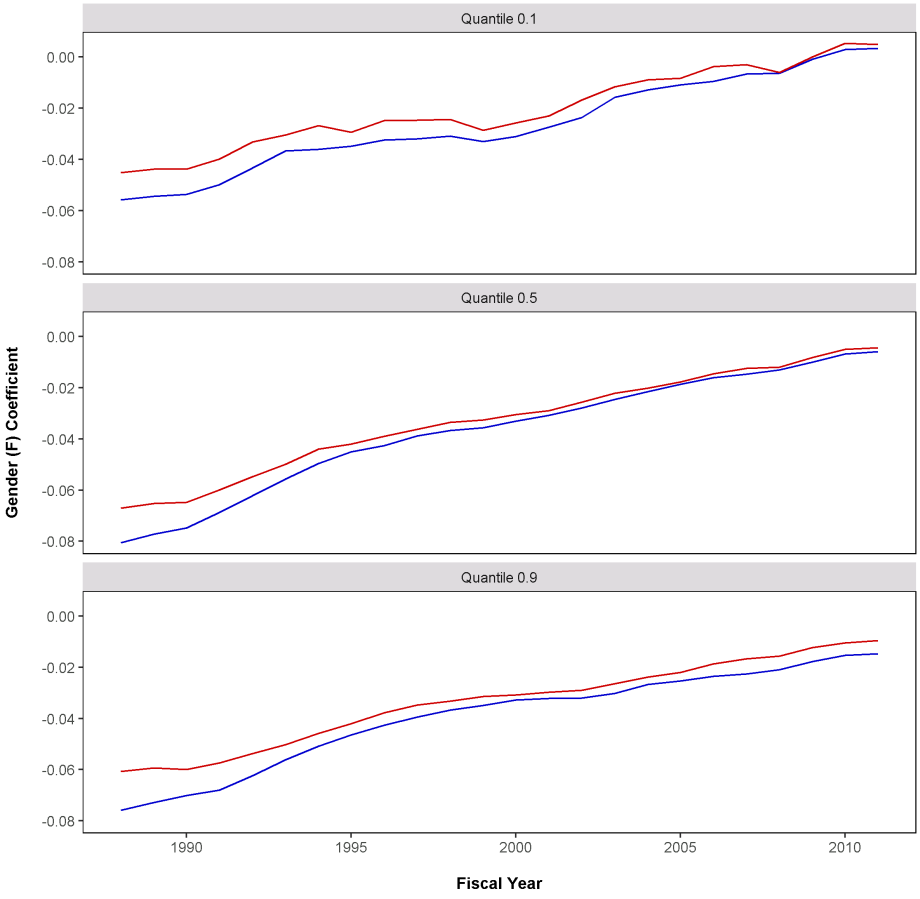


Figure 47: Pay disparity gender effect quantile estimates. Change over time. Upper line represents synthetic data, lower line authentic data.

8 The Rise of Grade in the U.S. Federal Government

This section uses results from a study of trends in human capital and pay within the U.S. federal government, conducted by the Human Capital Project at Duke University (Bolton and de Figueiredo, 2016). Each sub-section compares the fit of a model used in the study 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, C) and are identified as "FedScope."

8.1 Federal Wage Bill Decomposition

Figure 48 shows the annual change in the total federal employee wage bill, as reflected in the authentic OPM data set, 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.

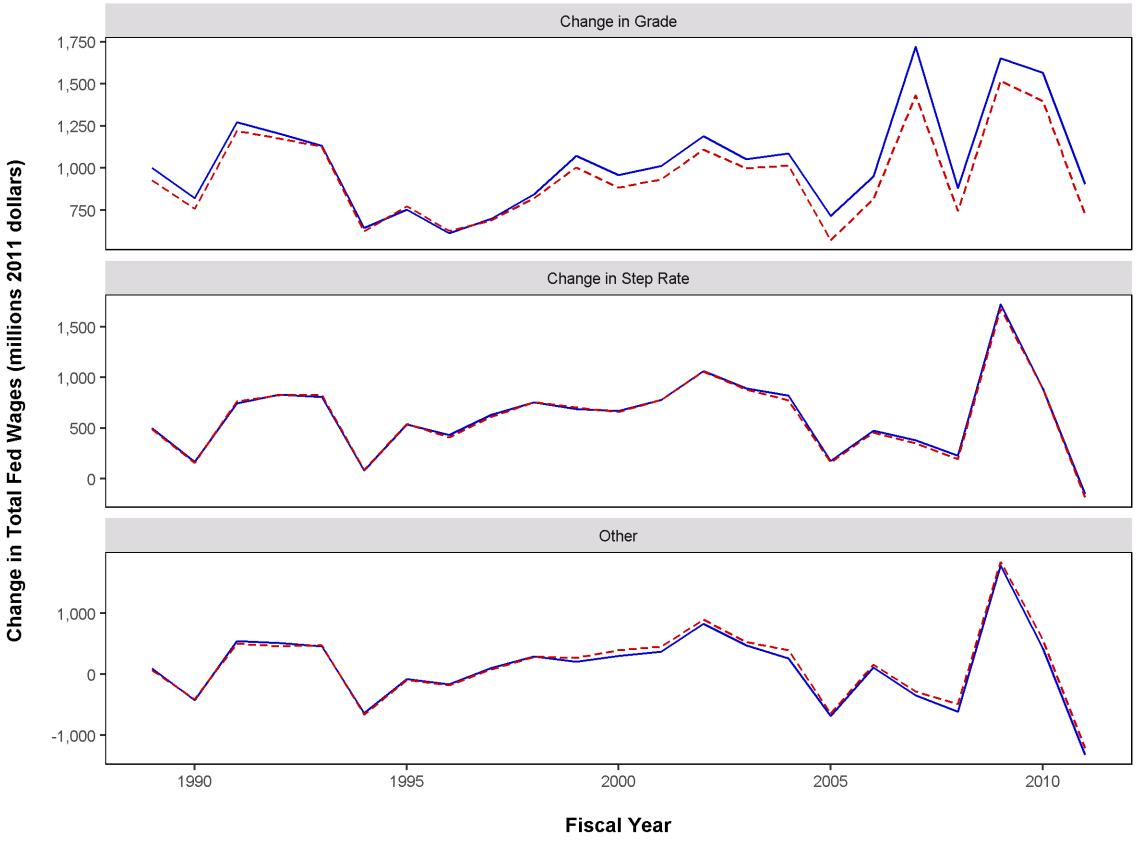


Figure 48: Change in U.S. federal government total wage bill. Millions of 2011 dollars by change in grade, change in step rate, and other changes. Fiscal years 1988 through 2011.

8.2 Change in GS Grade Distribution 2011 vs. 1988

GS pay plan, full-time observations represent approximately 75% of 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 49 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 and local change patterns as do the authentic data.

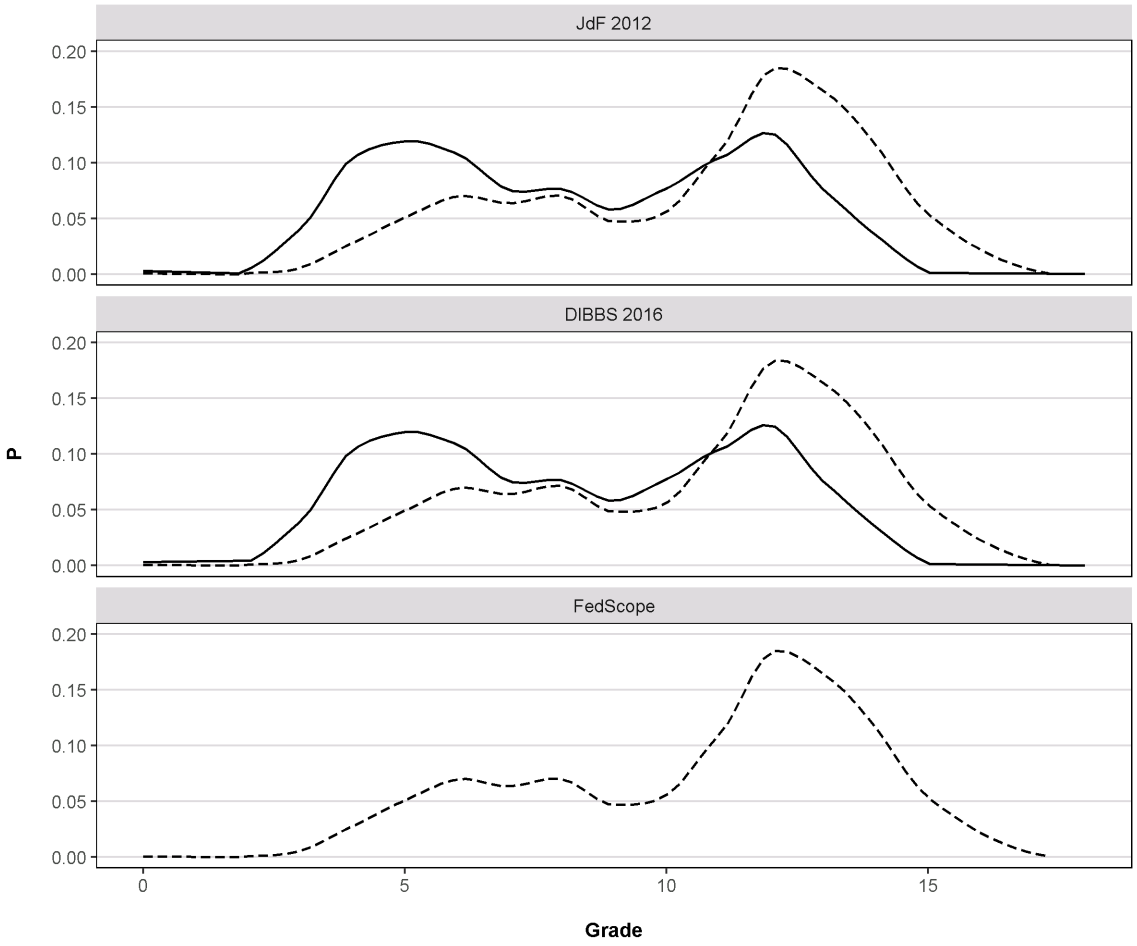


Figure 49: Change in GS grade distribution. Fiscal years 1988 (solid line) and 2011 (dashed line). Near identical distribution in synthetic and authentic data.

8.3 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 50 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.

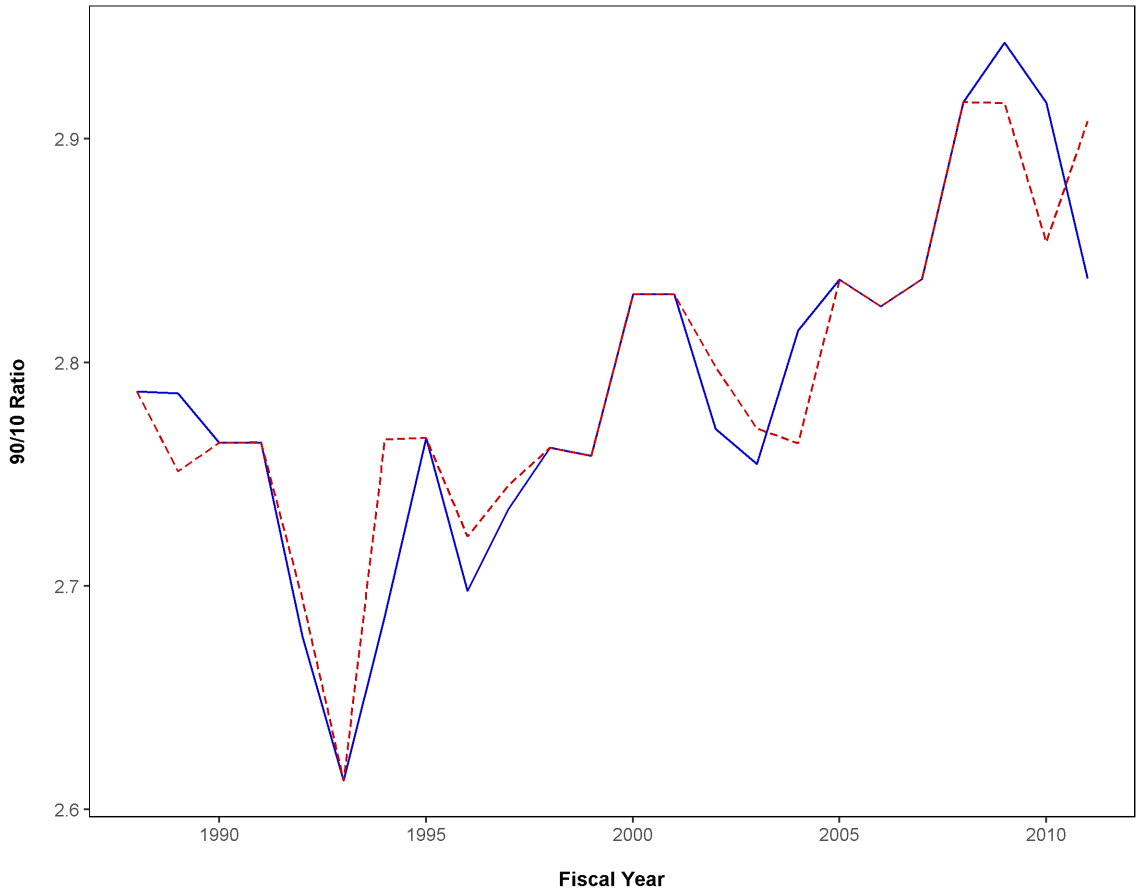


Figure 50: Ratio of 90th and 10th basic pay percentiles by year. GS pay plan, grade *leq* 15. Synthetic data dashed line, authentic data solid.

8.4 Basic Pay Quantile Regression

Ordinary least squares regression estimates the effect of independent, or predictor, variables on the expected value of a response. Of interest may be the association of time (fiscal year) with mean basic pay. Also of interest, when measuring increase in income, are the associations of time with key income quantiles, estimates of the effect of fiscal year on pay values below which a given proportion of observations are estimate to reside. Figure 51 plots, for pay plan GS, grade less than or equal to 15, the slopes estimated from the 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 is 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.

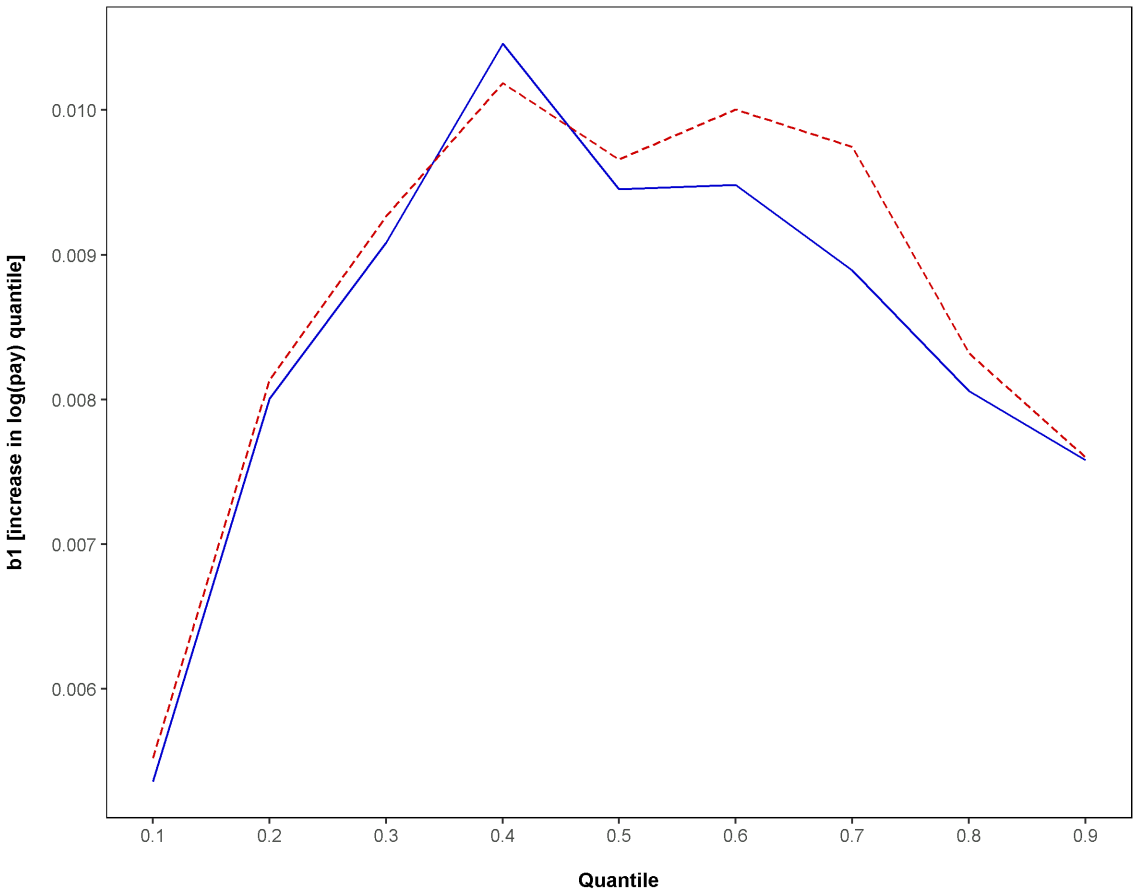


Figure 51: Coefficients (change per year) from quantile regression of log(pay) on year. GS pay plan, grade $\leq$ 15. 1988-2011. Synthetic data dashed line, authentic data solid.

8.5 Trend: Age of the U.S. Federal Employee

As a proxy for experience, employee age is an important independent variable in human capital research. Two aspects of age that Bolton and de Figueiredo measure are change, throughout their study period, in mean age of all employees and in mean age of first year employees. Figure 52 plots these means against fiscal year. Synthetic data are represented by a dashed line, authentic data by a solid line.

Observations: Mean age of all employees and first year employees increases throughout the study period, indicating an increasingly experienced workforce and apparent hiring of employees with increasing levels of experience. Although actual means of first year age are slightly underestimated in the synthetic data, overall and local authentic trends are accurately represented in the synthetic data.

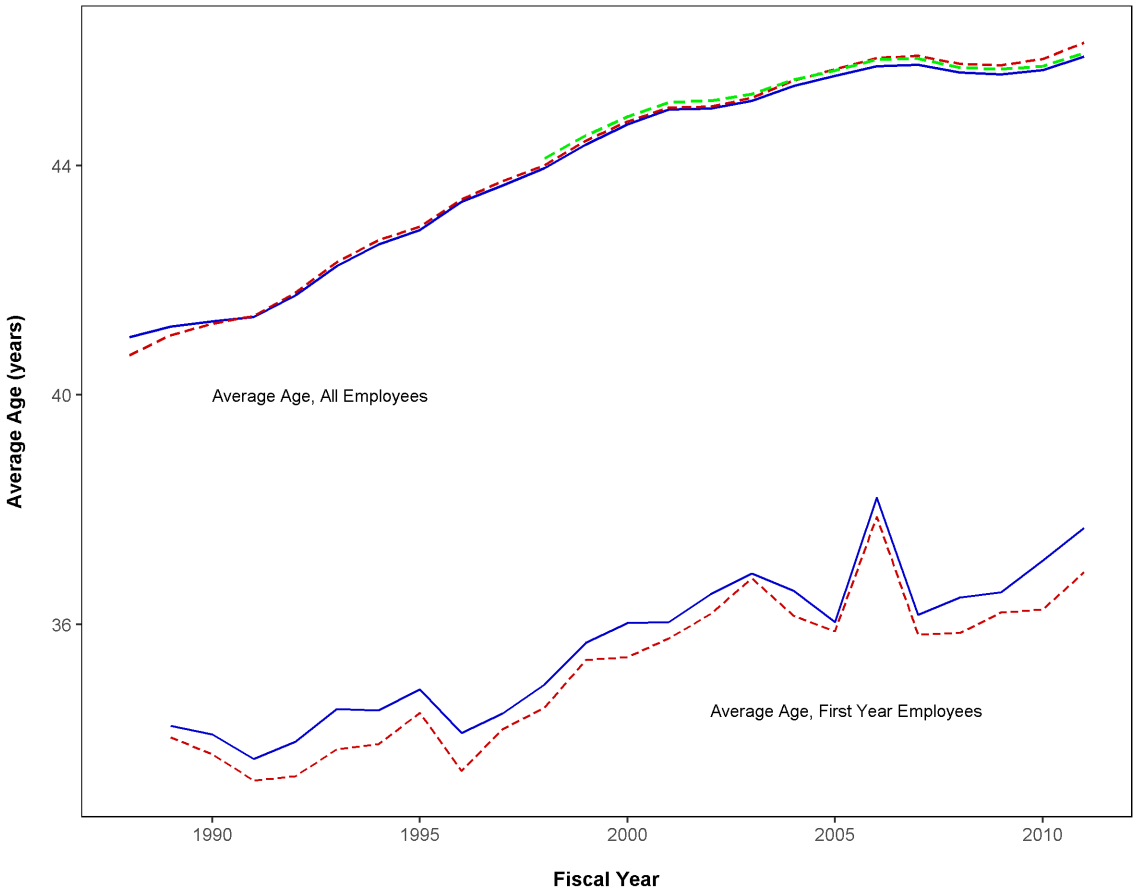


Figure 52: Change in mean age of all and first year federal employees. 1988-2011. Synthetic data dashed line, authentic data solid. Authentic trends accurately reflected in synthetic data.

8.6 Trend: Education Level of the U.S. Federal Employee

Employee education is an important independent variable in human capital research. Two aspects of education that Bolton and de Figueiredo measure are change, throughout their study period, in mean years of education for all employees and for first year employees. Figure 53 plots these means against fiscal year. Synthetic data are represented by a dashed line, authentic data by a solid line.

Observations: Mean years of education for all employees and first year employees increases throughout the study period, indicating an increasingly educated workforce. Although showing slight deviations from authentic annual means, means in the synthetic data accurately represented overall and local trends. The major reduction in 2002 is attributed to establishment of the Transportation Security Administration.

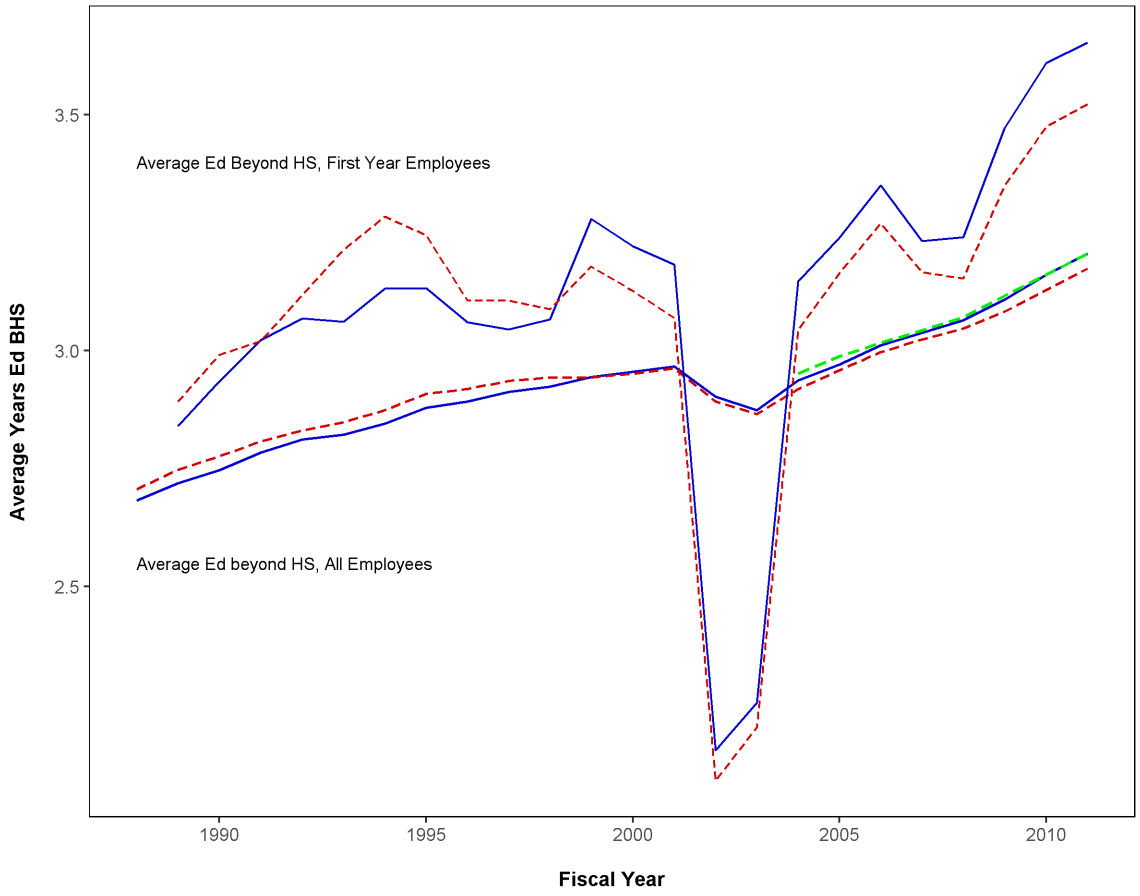


Figure 53: Change in mean years of education for all and first year federal employees. 1988-2011. Synthetic data dashed line, authentic data solid. Authentic trends accurately reflected in synthetic data.

8.7 Occupational Category Distribution

Bolton and de Figueiredo identify changes in the experience, education, and pay of federal employees over their study period and, additionally, a concurrent change in job classification, or occupational category. OPM classifies occupations as one of five types: professional, administrative, technical, clerical, other white collar, and blue collar. Figure 54 plots proportions of observations in the data by occupational category and fiscal year. Synthetic data are represented by a dashed line, authentic data by a solid line.

Observations: Throughout the study period, a significant increase in proportion professional and administrative positions occurs with corresponding reduction in clerical and blue collar positions, indicating major restructuring of occupations in the federal government. Reclassification may account for the increase in proportion technical occupations with concurrent decreases in proportion professional and clerical between 2000 and 2005. Although showing slight deviations from proportions in the authentic data, the synthetic data accurately represent overall and local trends, including corrections in the 2000-2005 period.

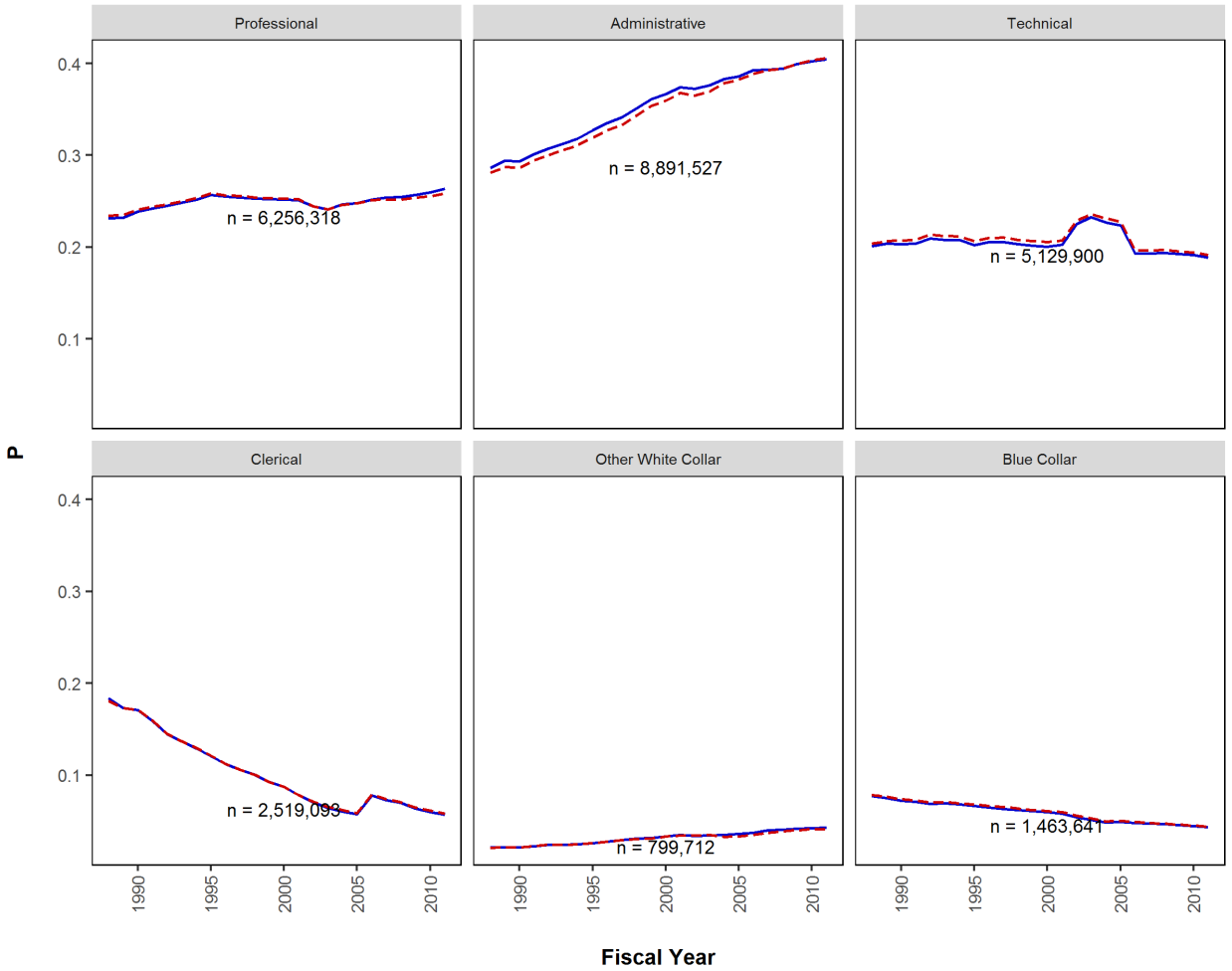


Figure 54: Change in structure of occupational categories in the U.S. federal government. 1988-2011. Synthetic data dashed line, authentic data solid. Authentic trends accurately reflected in synthetic data.

8.8 Job Switchers vs. Non-switchers, Age

The data supplied by OPM enable longitudinal career analysis. To study mobility within the federal government, Bolton and de Figueiredo measure mean difference in age (as a proxy for experience) between employees who who transition to a different occupational category and those who remain in their occupation, using a fixed effects regression model that controls for agency, occupation, and year. Figure 55 plots, for occupational categories P, A, T, C, and O, mean difference in age between employees who changed occupations in a given year and those who remain in their occupation that year.

Observations: All means are at or below zero, indicating a younger, less experienced sub-population of mobile workers in all occupational categories. This is reflected in both data sets. Although some discrepancies exist between means measured in the synthetic data and authentic data, the largest involve transition to category O, which are the lowest frequency. Interestingly, all means from synthetic data are greater than corresponding authentic means.

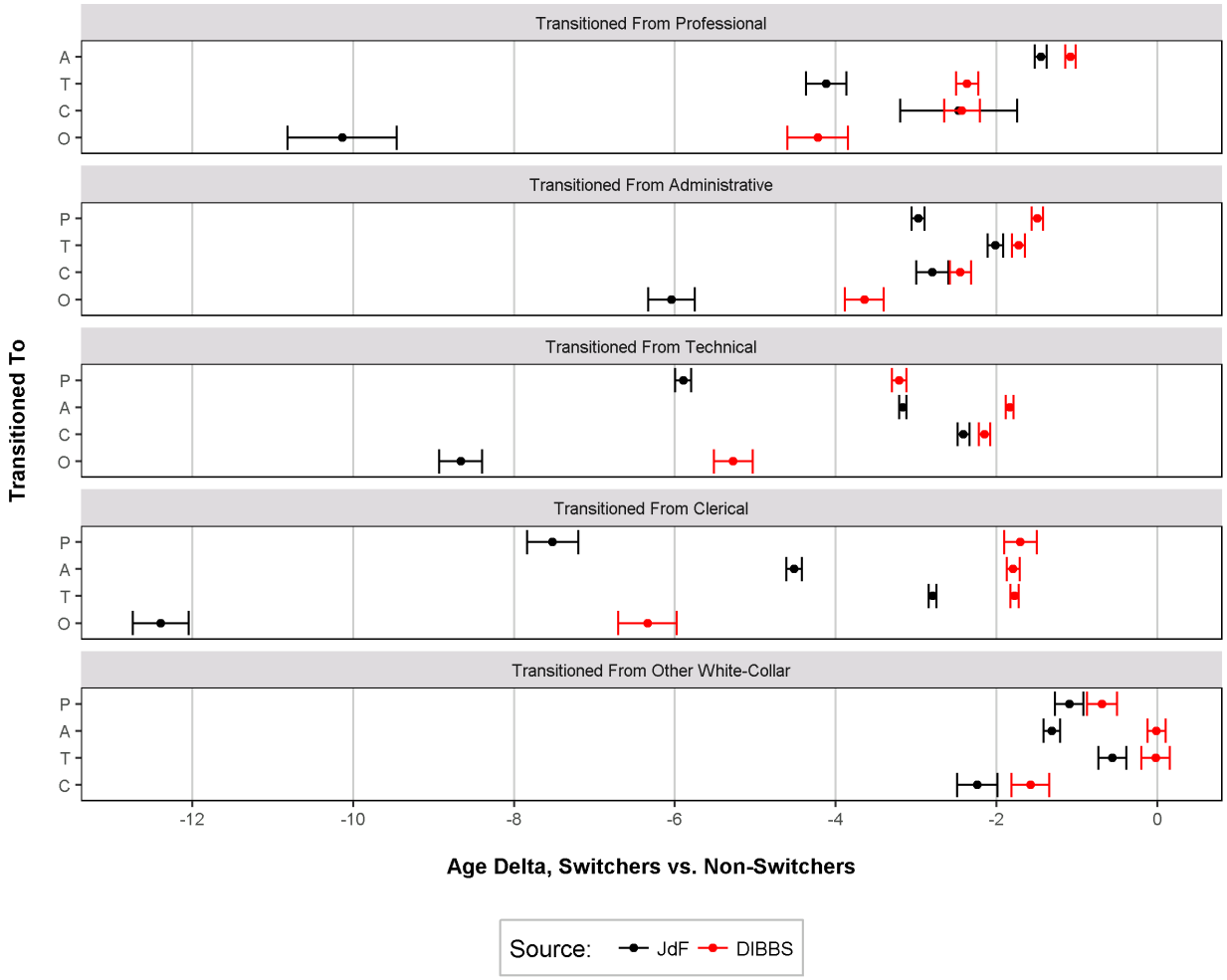


Figure 55: Mean difference in age between employees with change in occupational category and those who they join and who remained in their occupation. Non-positive means observed in both data sets.

8.9 Job Switchers vs. Non-switchers, Education

Continuing their study of mobility, Bolton and de Figueiredo measure mean difference in years of education between employees who who transition to a different occupational category and those who remain in their occupation, using a fixed effects regression model that controls for agency, occupation, and year. Figure 56 plots, for occupational categories P, A, T, C, and O, mean difference in education (years) between employees who changed occupations in a given year and those who remain in their occupation that year.

Observations: Mean difference in years of education appears to be associated with occupational category transitioned from, with nearly all means for categories T, C, and O being positive and all for categories P and A being negative. This indicates that mobile employees in technical, clerical, and other white collar occupations tend to have higher levels of education than those who they join (employees who remained) in their new occupation. All but one mean from the P and A categories are negative, indicating a less educated group who transition from professional and administrative occupations, when compared to their new coworkers. Bolton and de Figueiredo offer compelling explanations for these patterns that the reader may find interesting. Although with different specific values, means from the synthetic data inform on the important general findings from the authentic data.

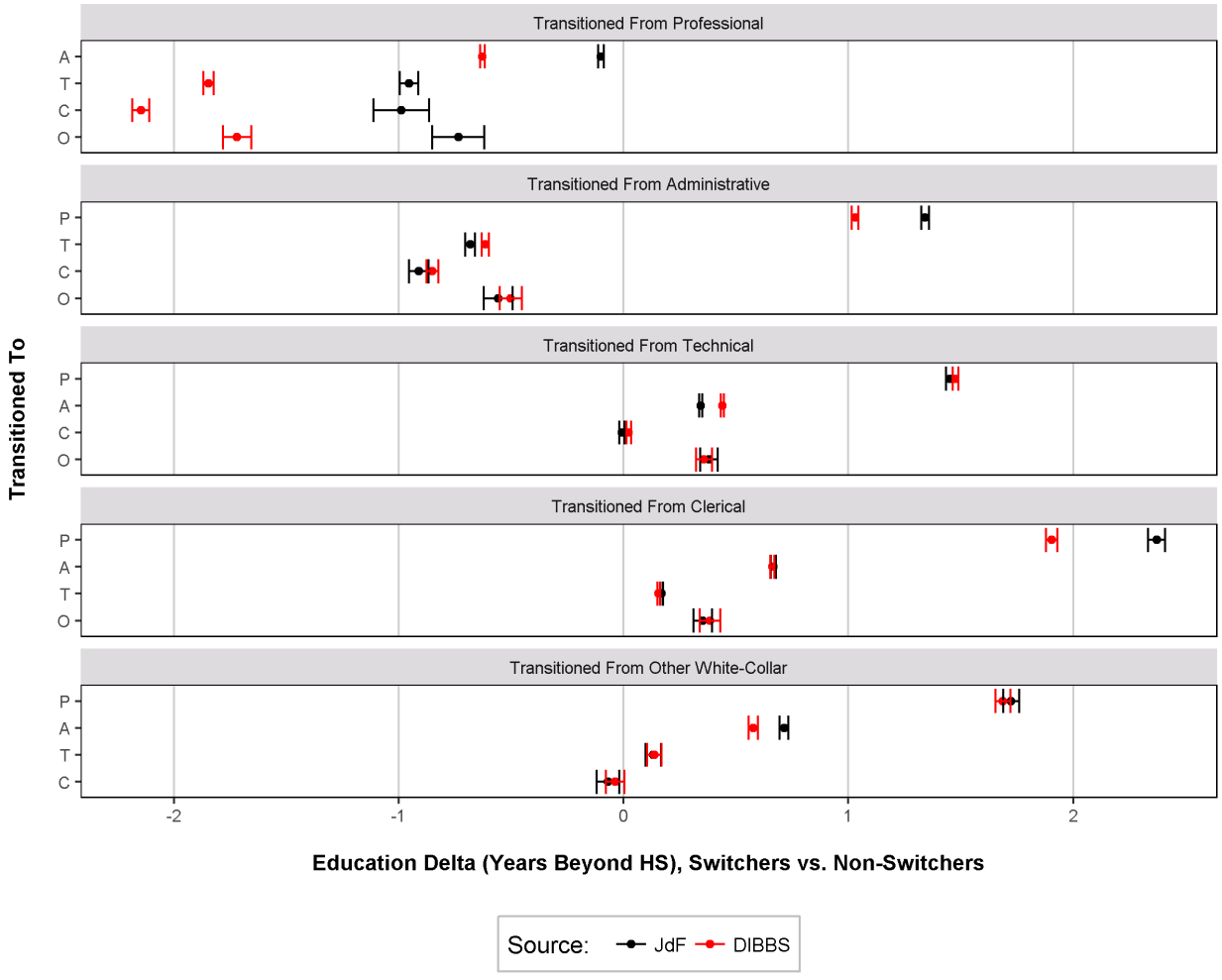


Figure 56: Mean difference in years of education between employees with change in occupational category and those who they join and who remained in their occupation. Positive means associated with categories T, C, and O, negative means associated with categories P and A.

9 Logistic Regression Promotion Model

In addition to studying disparities in pay, human capital researchers are interested in measuring possible disparities in promotion that are associated with gender or race. For the federal employee, OPM defines promotion as a positive change in grade and, since their data set has one observation per employee per year, proportion employees promoted per year within a given category (of, say, sex, race, grade, and age) can be computed as the ratio of the number of observations with increase in grade, by employee, since their most recent work year to the total number of employee-year observations in the category. Limiting data to full-time observations within the GS pay plan, the logistic regression model

$$\hat{\boldsymbol{p}} = f(\hat{\boldsymbol{\beta}}_{sex} sex + \hat{\boldsymbol{\beta}}_{race} race + \hat{\boldsymbol{\beta}}_{grade} grade + \hat{\boldsymbol{\beta}}_{age} age) \tag{3}$$

estimates the proportion of employees promoted per year within each distinct group of sex, race, grade and age.[8] [9] [10] Figures 57 through 59 plot observed proportion employees promoted per year as a function of age, given sex, race, and grade for grades 3 through 7, grades 8 through 12, and grades 13 through 17, respectively. Figures 57 through 59 plot proportions estimated from model (3) fit to observed proportions for grades 3 through 7, grades 8 through 12, and grades 13 through 17, respectively.

Observations:

[8] The GS pay plan has numeric grades, which are convenient for identifying increases. Also, GS, full-time observations account for approximately 75% of observations supplied by OPM.

[9] Since observations are restricted to those for the GS pay plan, promotions to or from other pay plans are excluded here.

[10] Age is used as a proxy for work experience.

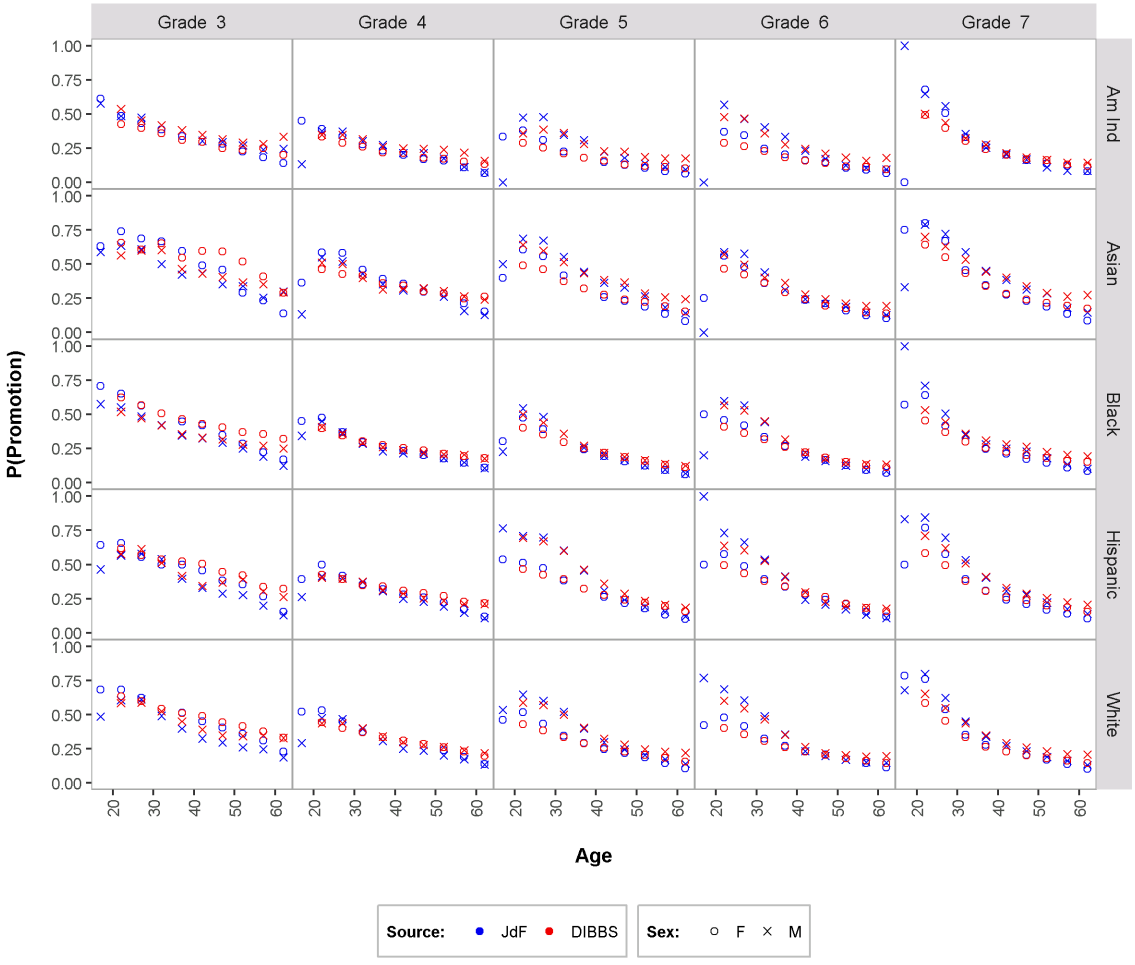


Figure 57: Observed proportion employees promoted per year as a function of age (experience) given sex, race, and grade. 1989-2011. GS pay plan, grades 3 through 7.

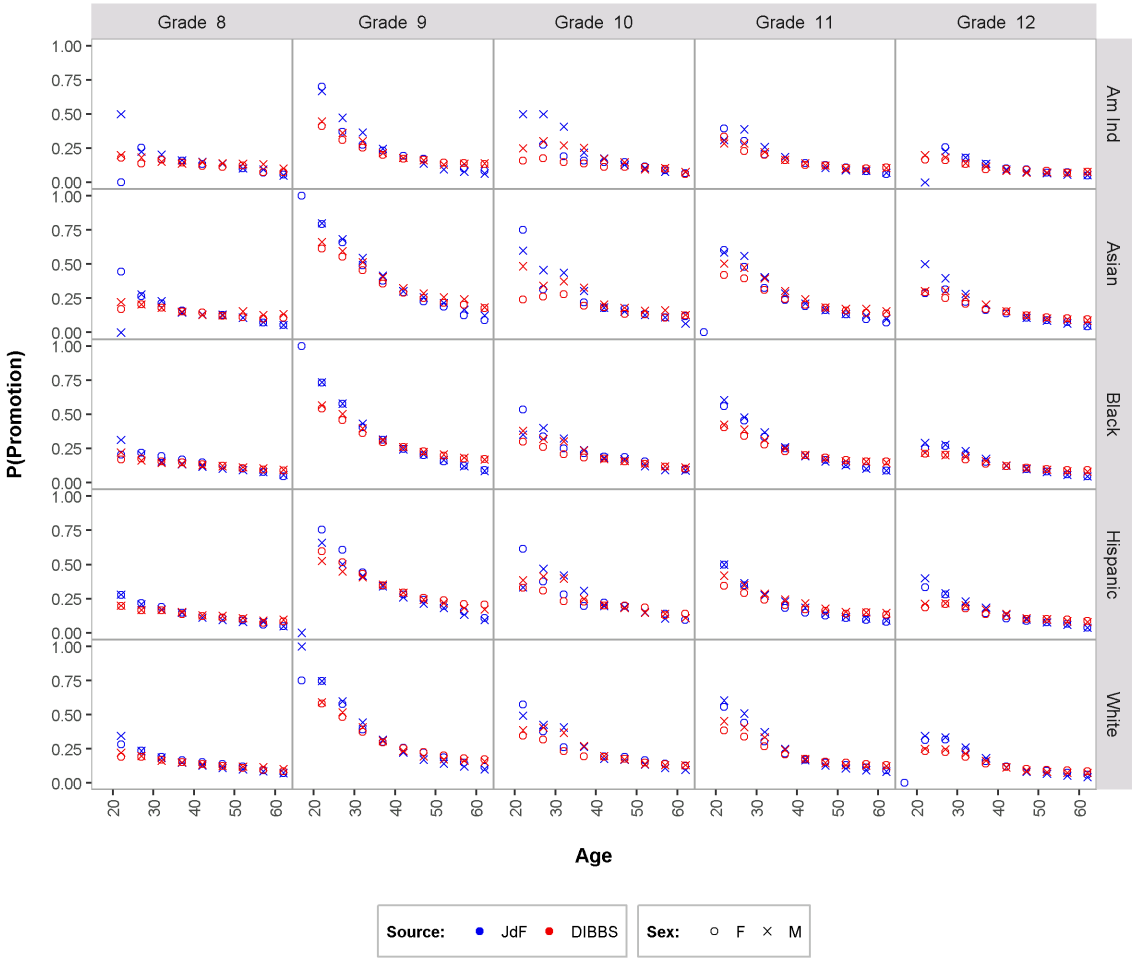


Figure 58: Observed proportion employees promoted per year as a function of age (experience) given sex, race, and grade. 1989-2011. GS pay plan, grades 8 through 12.

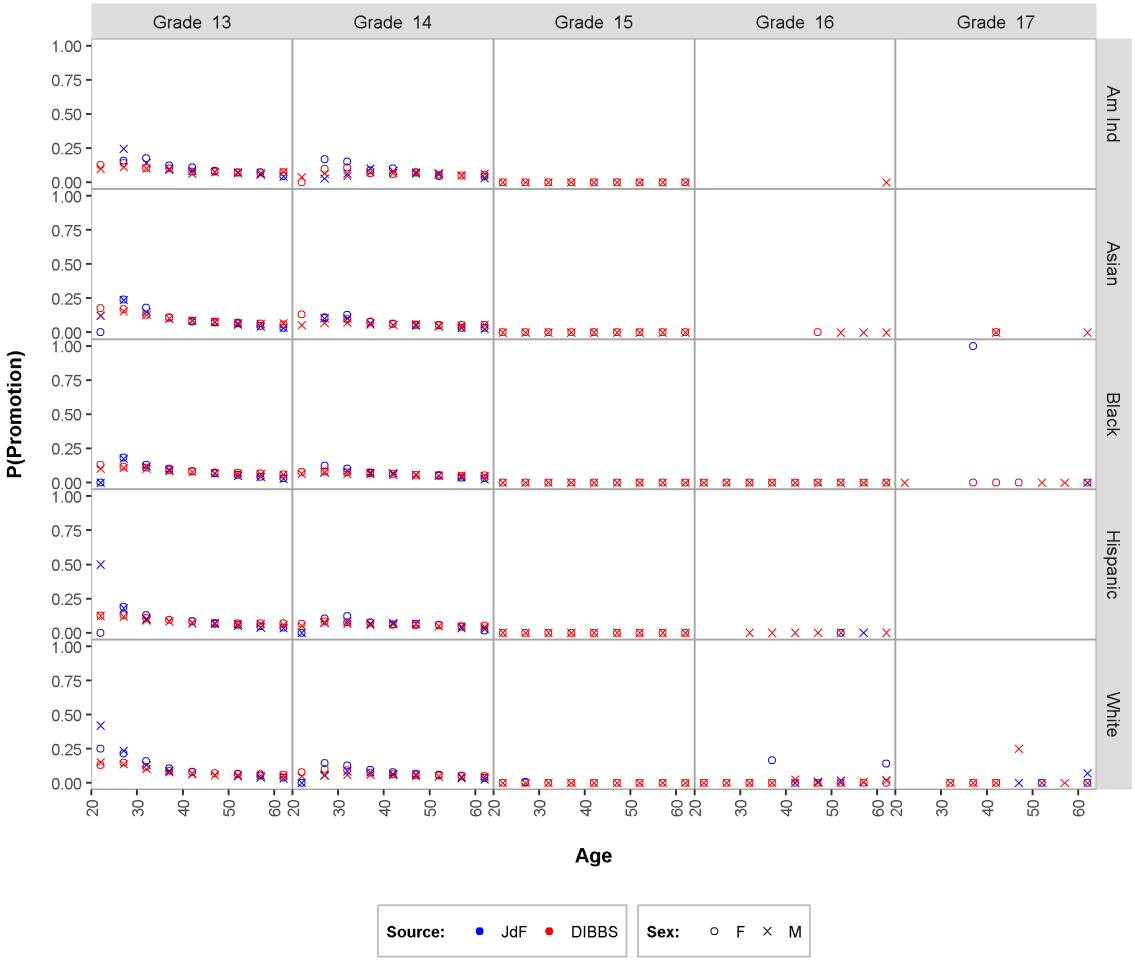


Figure 59: Observed proportion employees promoted per year as a function of age (experience) given sex, race, and grade. 1989-2011. GS pay plan, grades 13 through 17.

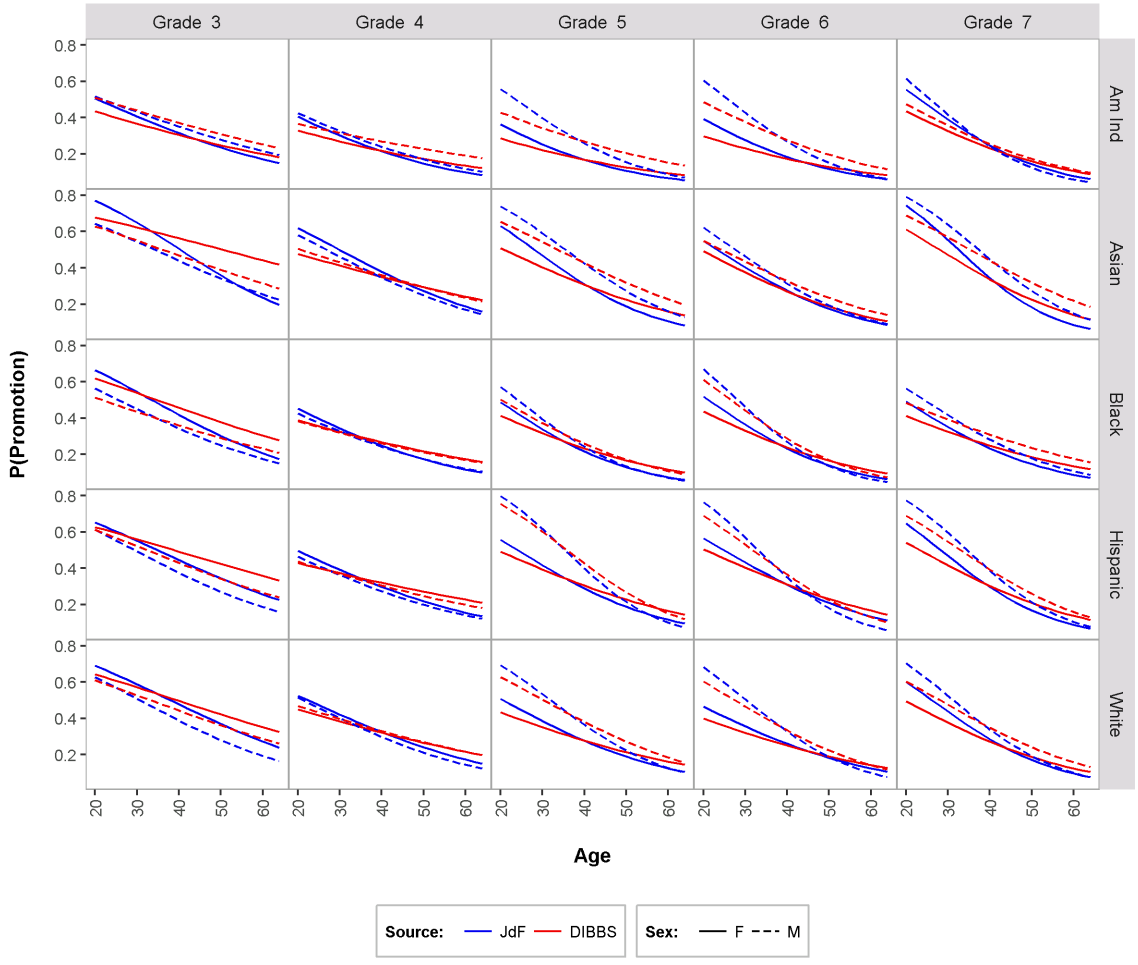


Figure 60: Logistic regression estimates of proportion employees promoted per year as a function of age (experience) given sex, race, and grade. 1989-2011. GS pay plan, grades 3 through 7.

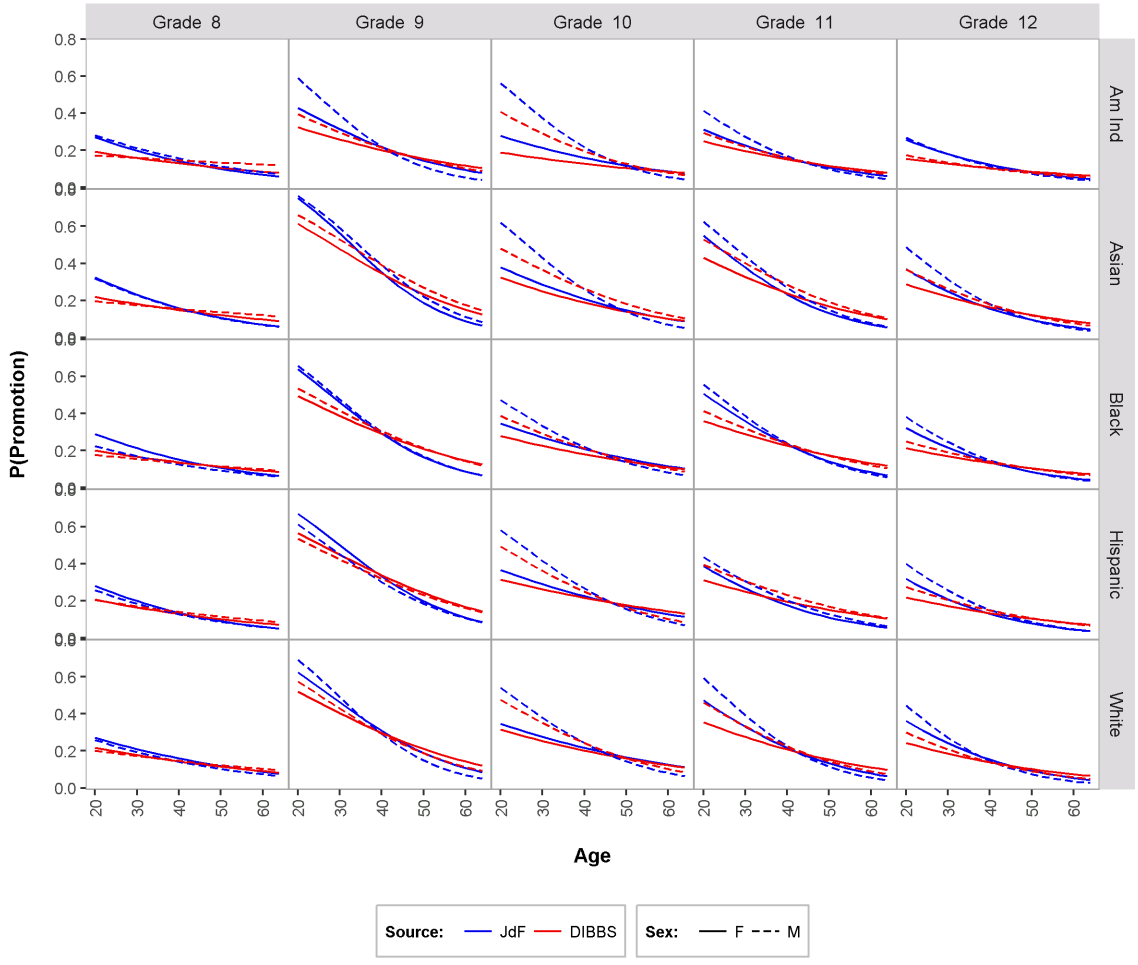


Figure 61: Logistic regression estimates of proportion employees promoted per year as a function of age (experience) given sex, race, and grade. 1989-2011. GS pay plan, grades 8 through 12.

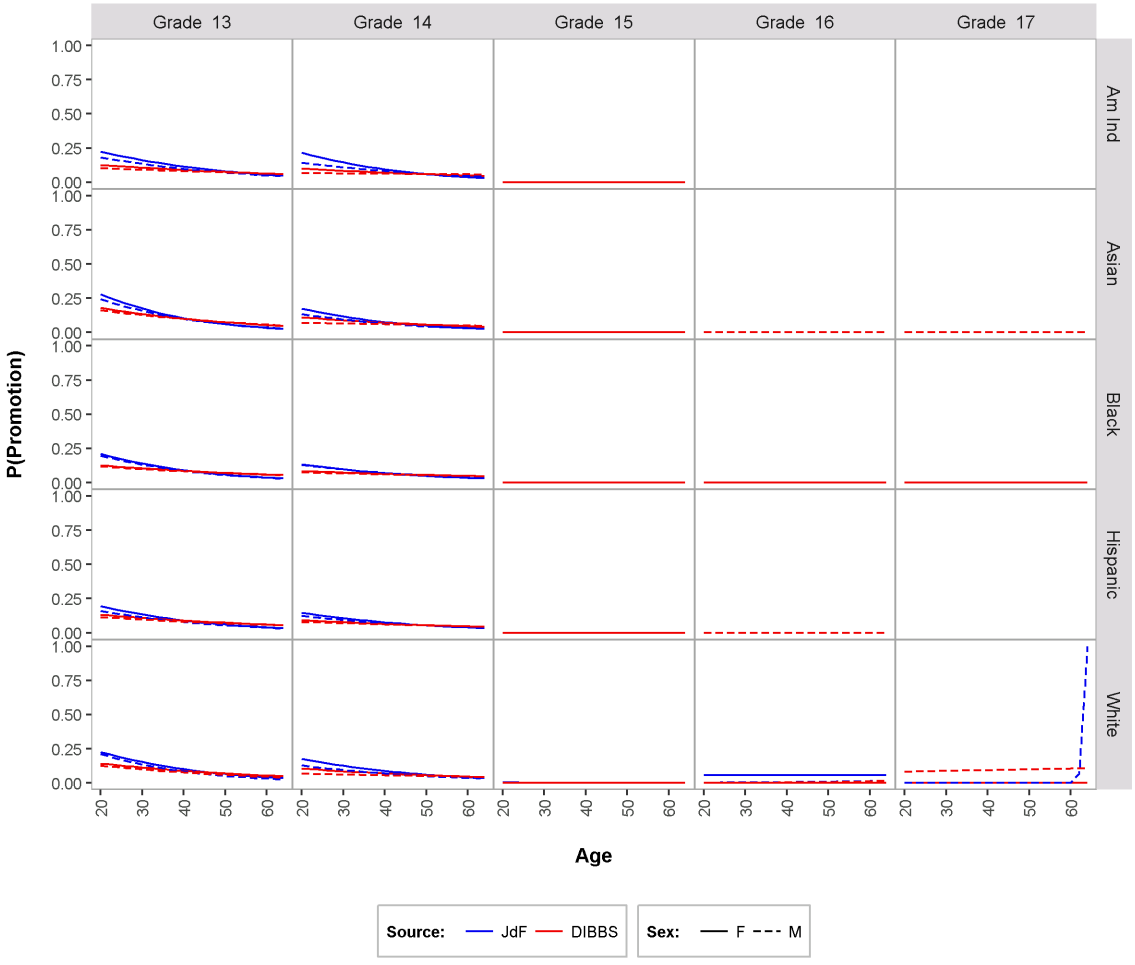


Figure 62: Logistic regression estimates of proportion employees promoted per year as a function of age (experience) given sex, race, and grade. 1989-2011. GS pay plan, grades 13 through 17.

10 Longitudinal Employee Careers

Relationships of human capital factors such as experience, education, and demographics to agency and occupation employed in, along with patterns of change through time, are important areas of study. To assess the U.S. federal employee's readiness to develop and execute policy, along with associated costs, researchers model promotion, changes in distribution of professional class and grade within subpopulations, and patterns of mobility between agencies and occupations. Some models include ancillary data to compare human capital profiles of public sector employees to compatible workers in the private sector. Some measure interaction between events within and without official governmental programs. For the synthetic data to be useful in these contexts, longitudinal relationships must be maintained and, in fact, it is with agency and year that the synthesis algorithm begins. As seen earlier, in table 1, the synthetic data have slightly more observations than the authentic data, but with slightly fewer employees and contain no new year, agency, occupation combinations. Figure 63 plots the ratio, by year, of synthetic observations to total authentic observations. The dotted line is the ratio of total synthetic to total authentic observations. The solid line is the accumulated proportion, by year, of authentic observations within agency that appear in the synthetic data. These ratios deduct for under-representation and omit over-represented observations by agency. Note that duplicate year, employee ID, and agency in the authentic data are reduced to a single instance.

Observations: There exists an additional, approximate 1.25% total synthetic records per year, but observation frequency proportions within year and agency remain high at approximately 98% of authentic counts. Although cross-sectional, since annual totals are independent, it appears that the volume of synthetic observations by agency in each year is consistent with that needed to accurately represent careers in the authentic data. A slight, gradual increase in proportion observations within agency by year is apparent.

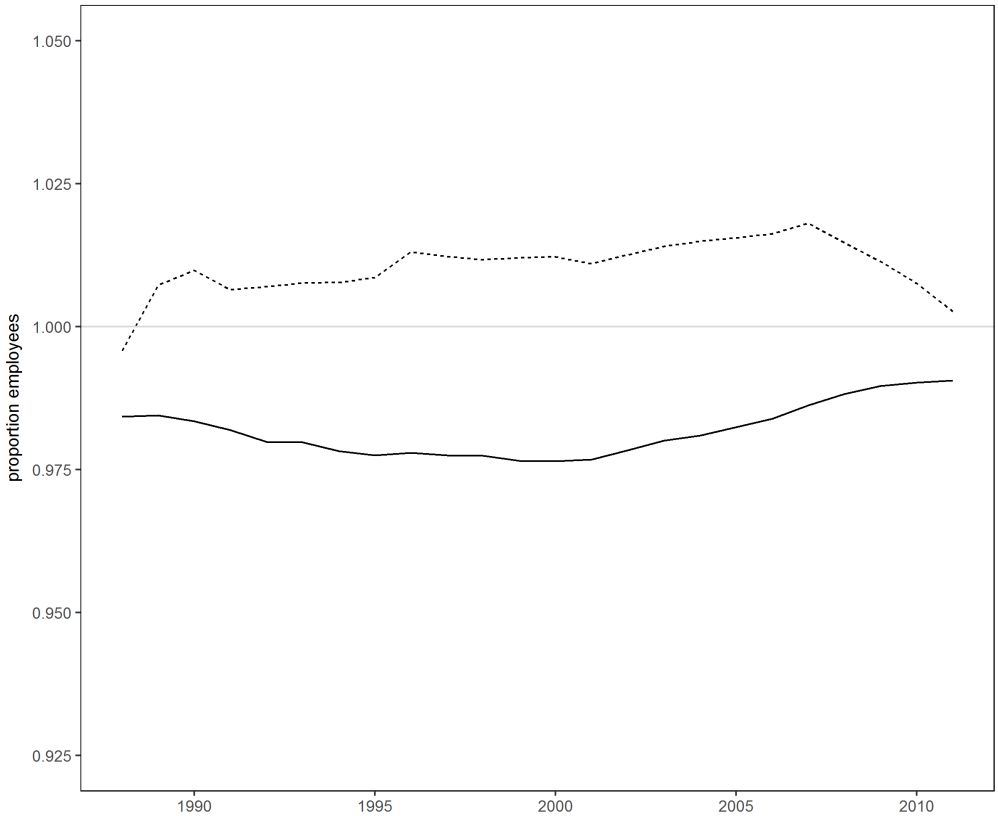


Figure 63: Annual ratios of synthetic to authentic observations. Dotted line is the ratio of total synthetic to total authentic observations. Solid line is the accumulated proportion, by year, of authentic observations within agency that appear in the synthetic data.

10.1 Careers by Consecutive Year Agencies

Explanation

Observations: Line corresponding to n=1 is identical to proportion authentic year, agency observations in synthetic data in figure 63 (solid line). This is because single year career segments are equivalent to cross-sectional proportions by year.

Patterns of line for n years influences shape of lines for subsequent n. *Explain*

Proportions above 0.88, even for 24 year careers.

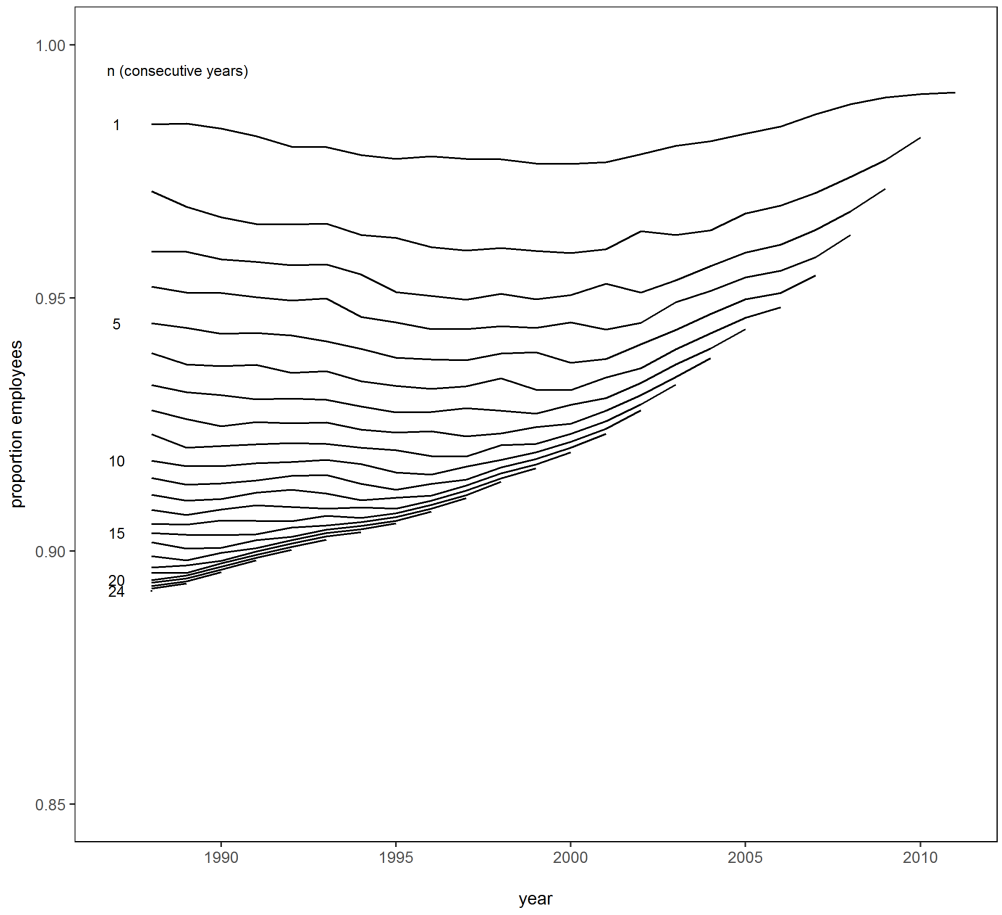


Figure 64: Proportion of authentic careers represented in synthetic data. Careers based on agencies employed in during n consecutive years. One line for each n from 1 to 24. Points on line represent proportion for n-year segment beginning in corresponding year on x-axis.

10.2 Careers by Consecutive Year Agency and Occupation

Explanation

Observations: Proportions generally below those for careers based solely on agency.

Lines generally more flat than those for agency alone, particularly for small n.

Proportions above 0.65, indicating possible limitation to accurate modeling of careers, particularly lengthy ones.

Curious ridge.

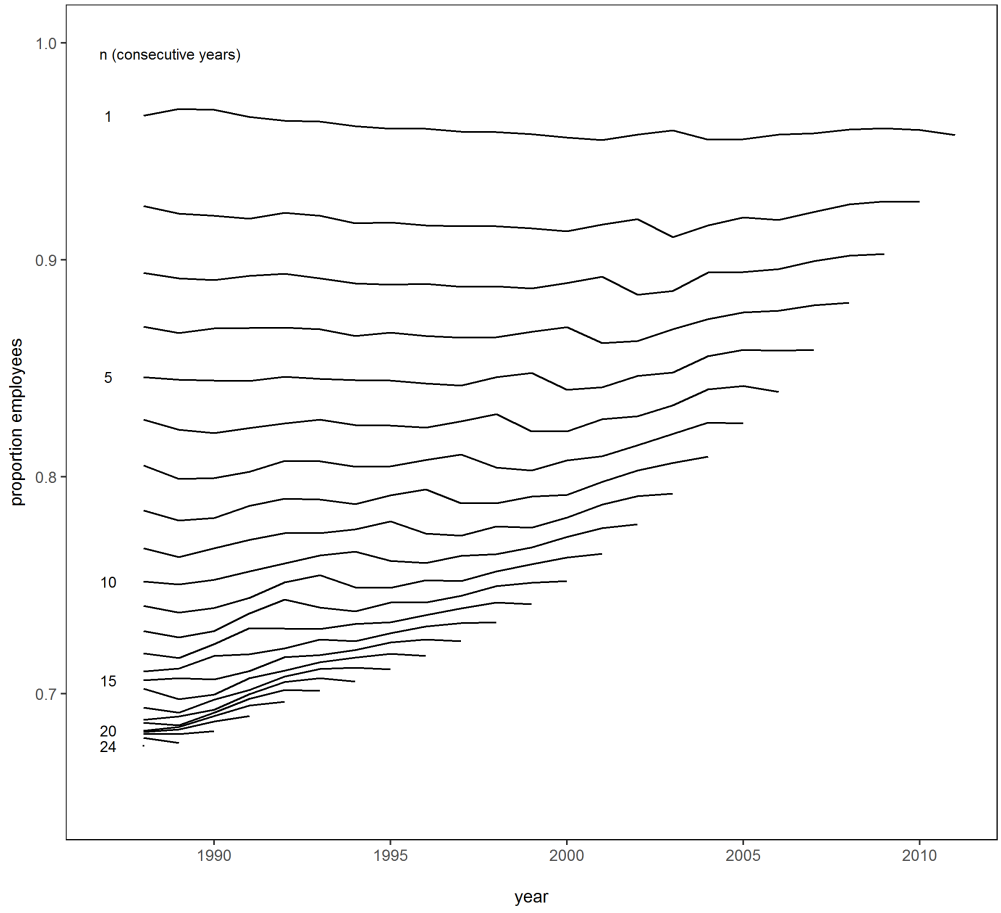


Figure 65: Proportion of authentic careers represented in synthetic data. Careers based on agencies and occupations employed in during n consecutive years. One line for each n from 1 to 24. Points on line represent proportion for n-year segment beginning in corresponding year on x-axis.

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, A. Guide to Data Standards. URL https://catalog.data.gov/dataset/guide-to-data-standards-gds.

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

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