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 6, 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 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.

1.1 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.

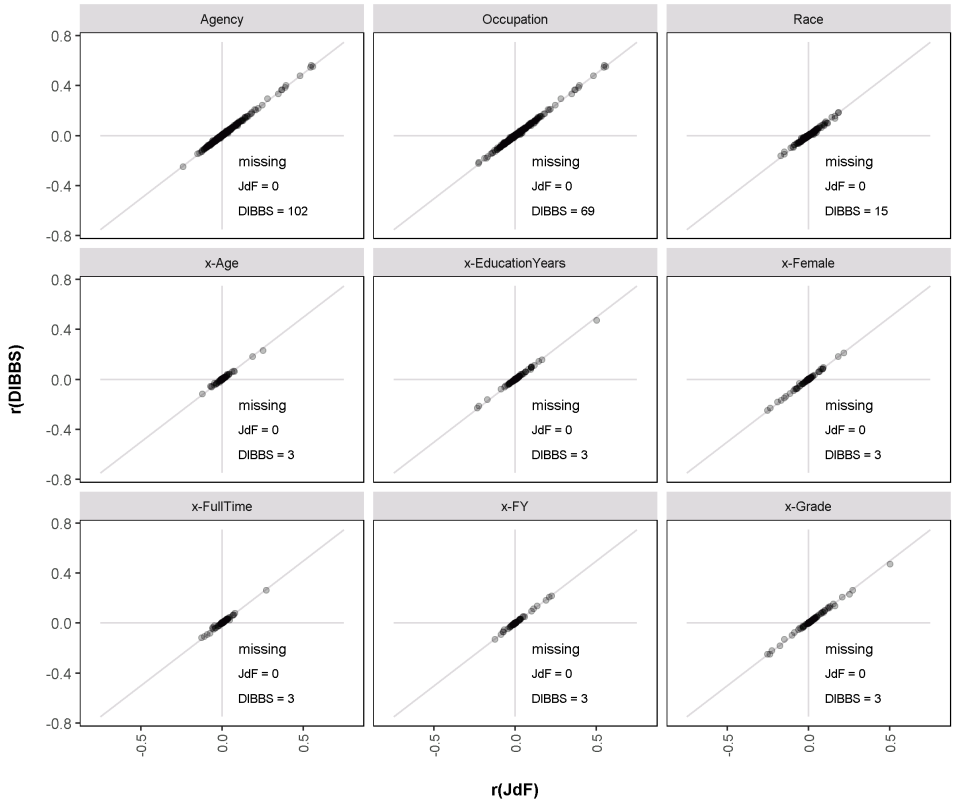


Figure 1: Two variable correlations of corresponding levels of synthetic and authentic data. Synthetic level correlation on y-axis, corresponding authentic correlation on x-axis.

1.2 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.

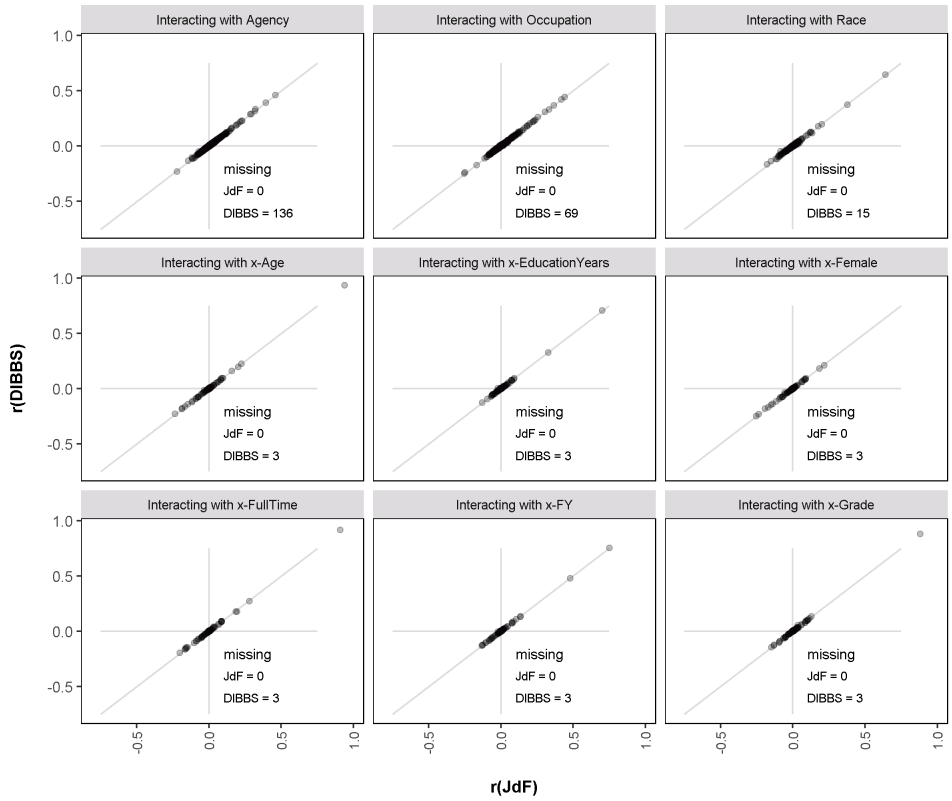
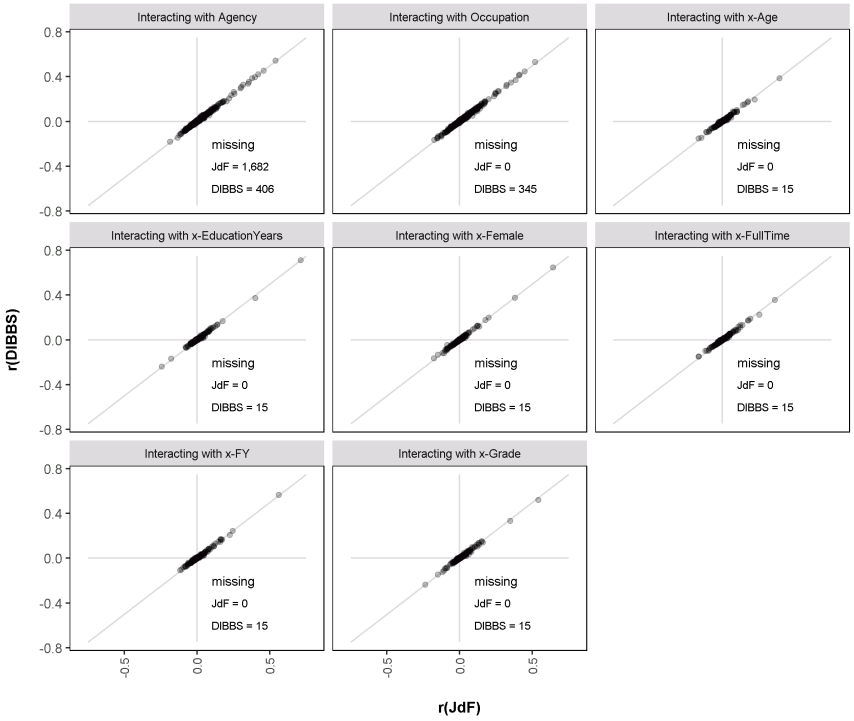
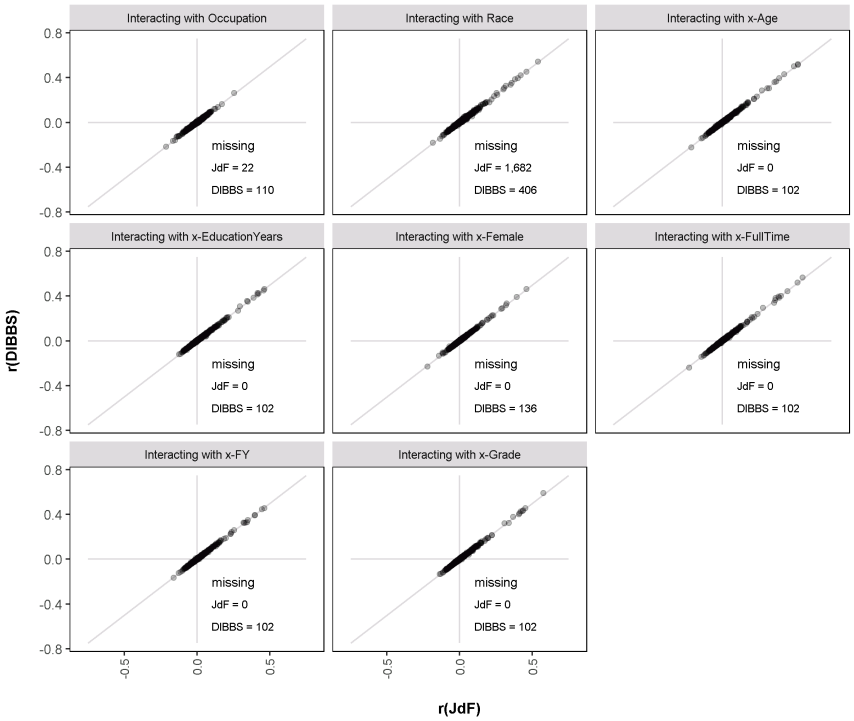


Figure 2: Correlation of primary variables with two variable interactions. Variable set one, involving sex. Synthetic level correlation on y-axis, corresponding authentic correlation on x-axis.

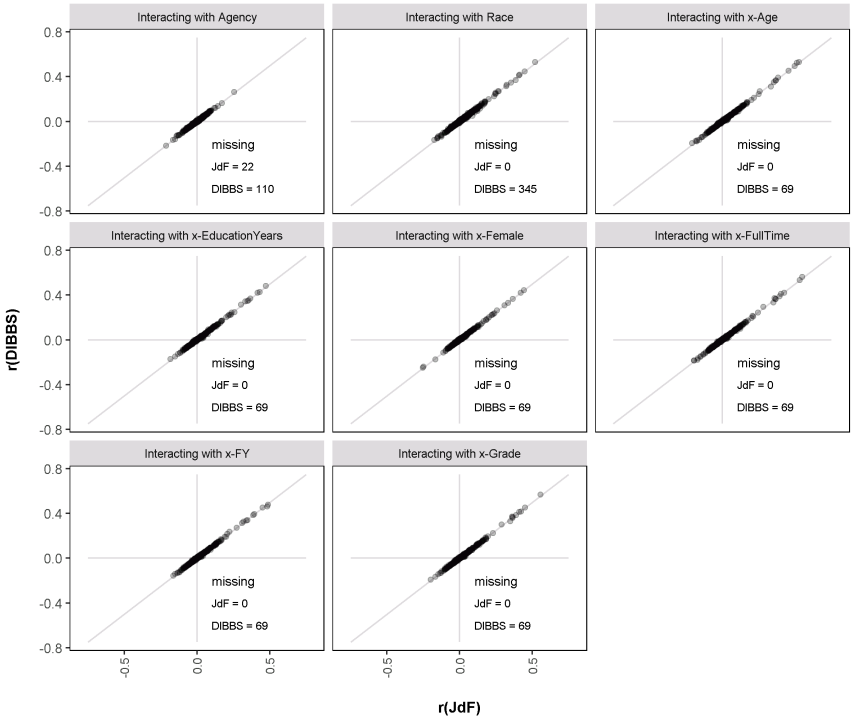


(a) Correlations involving race

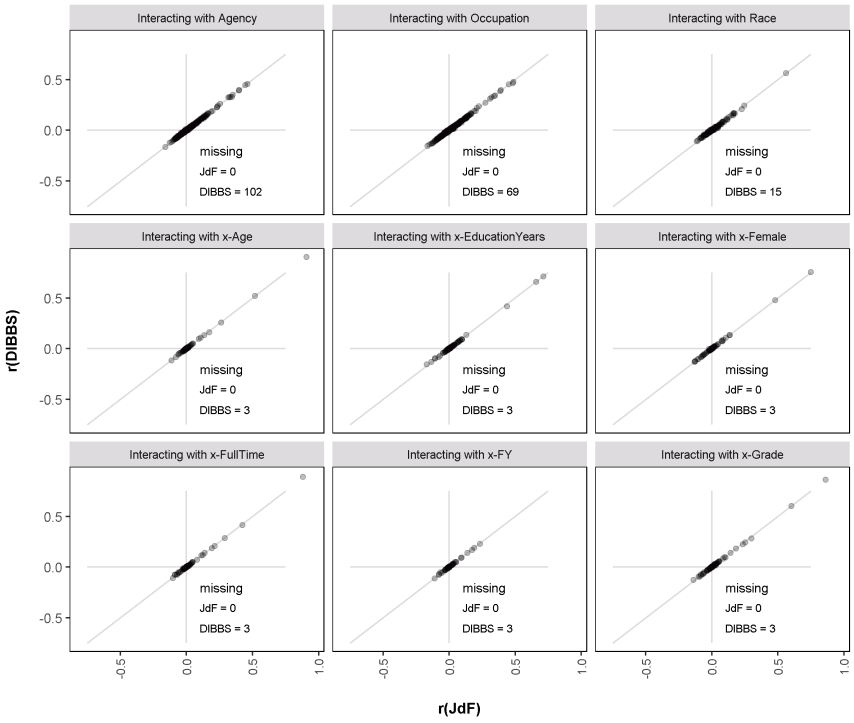


(b) Correlations involving agency

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.

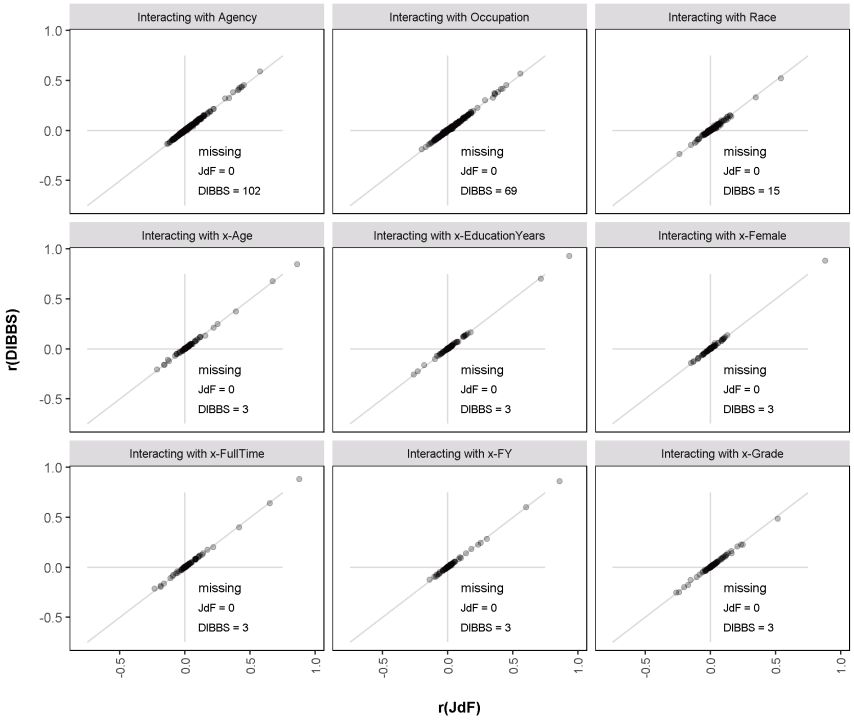


(a) Correlations involving occupation

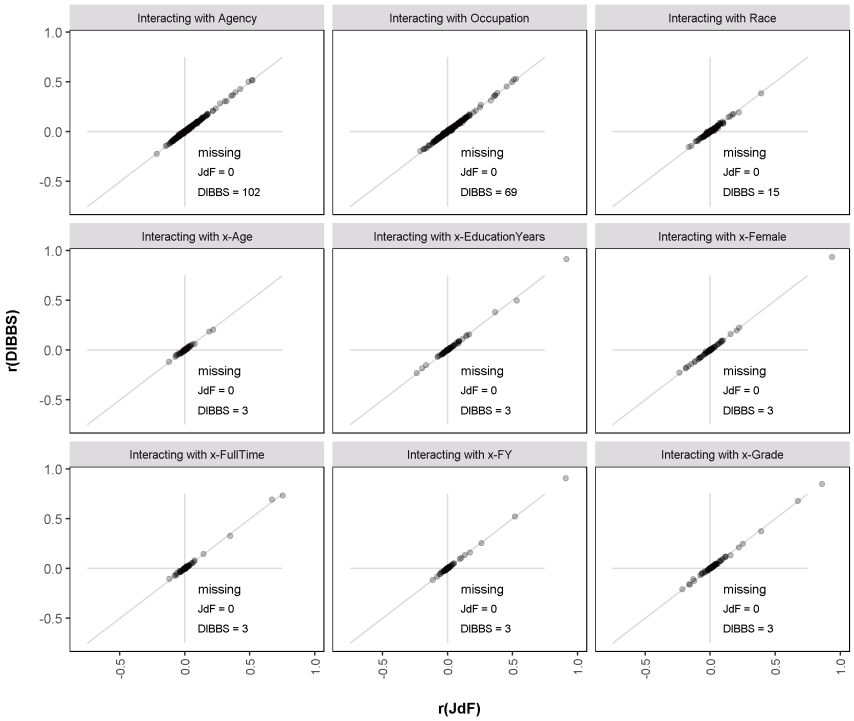


(b) Correlations involving fiscal year

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.

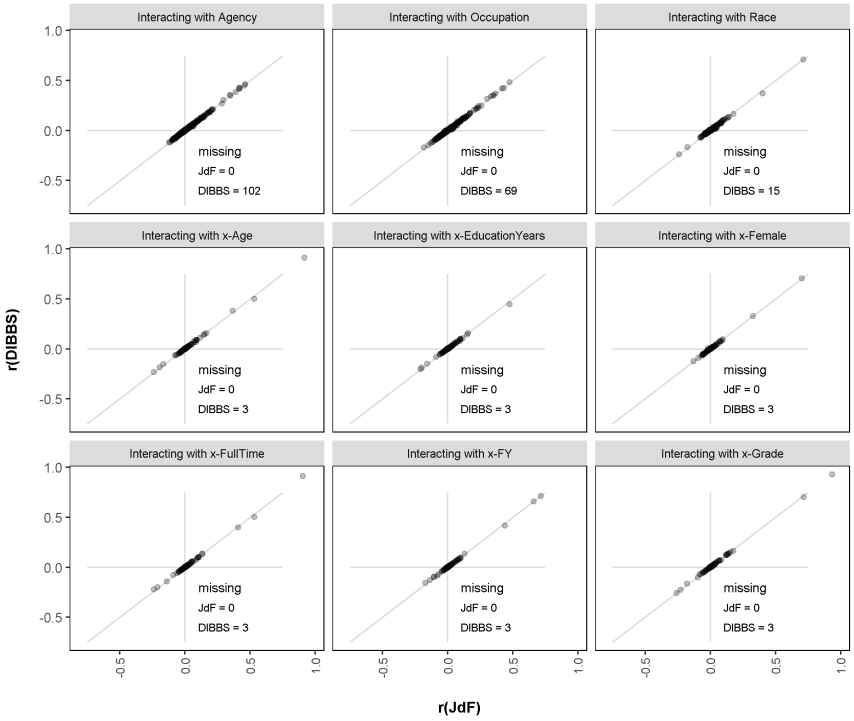


(a) Correlations involving grade

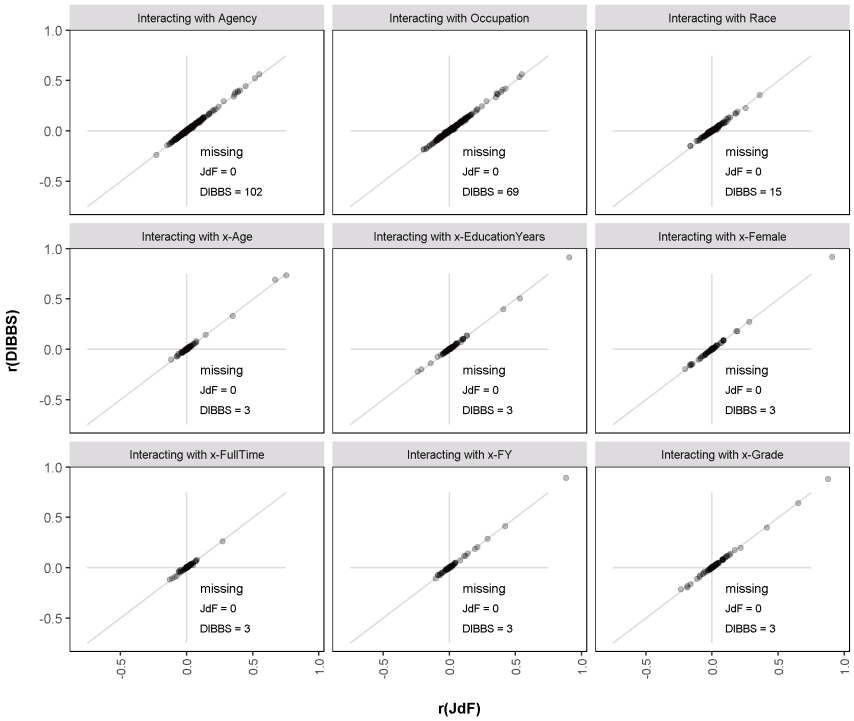


(b) Correlations involving age

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.

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

Figures 7 and 8 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.

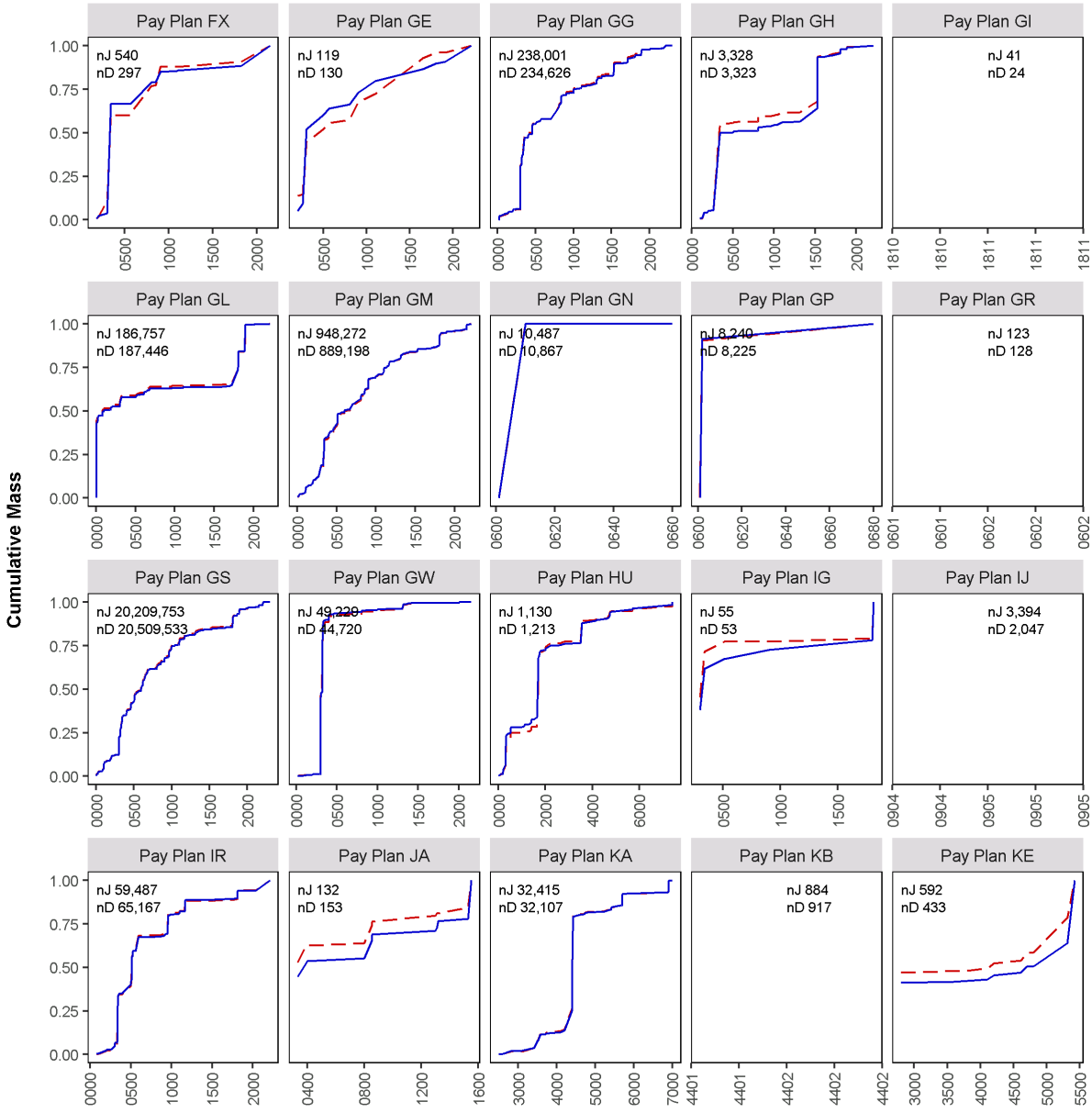


Figure 7: Cumulative mass by occupation within pay plan. Pay plan set one. Synthetic data represented by dashed line, authentic by solid line.

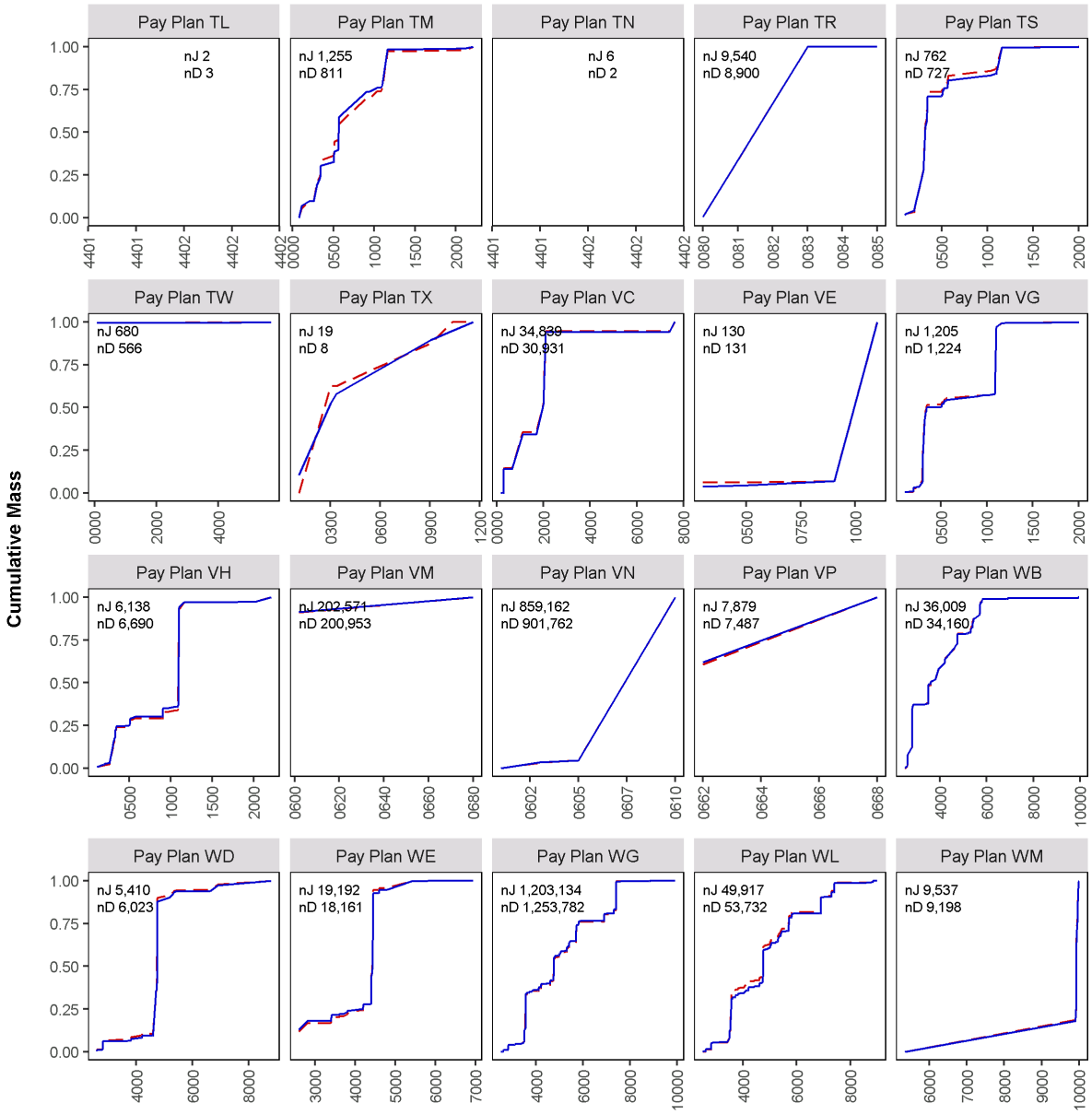


Figure 8: Cumulative mass by occupation within pay plan. Pay plan set 2. Synthetic data represented by dashed line, authentic by solid line.

3 Distribution of Basic Pay

3.1 Distribution of Basic Pay by Agency

Figure 9 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.

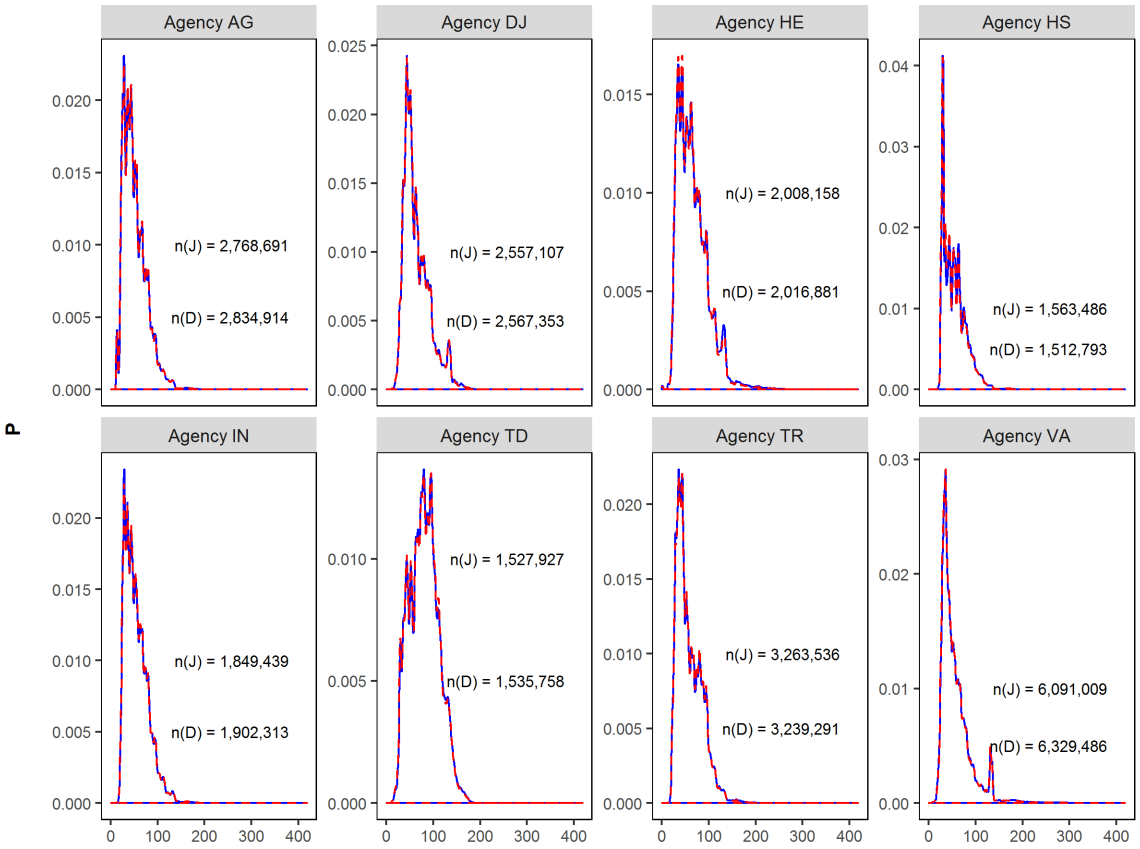


Figure 9: Basic pay marginal distribution for top eight agencies. Dashed line for synthetic data, solid line for authentic.

3.2 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 10 through 17 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.

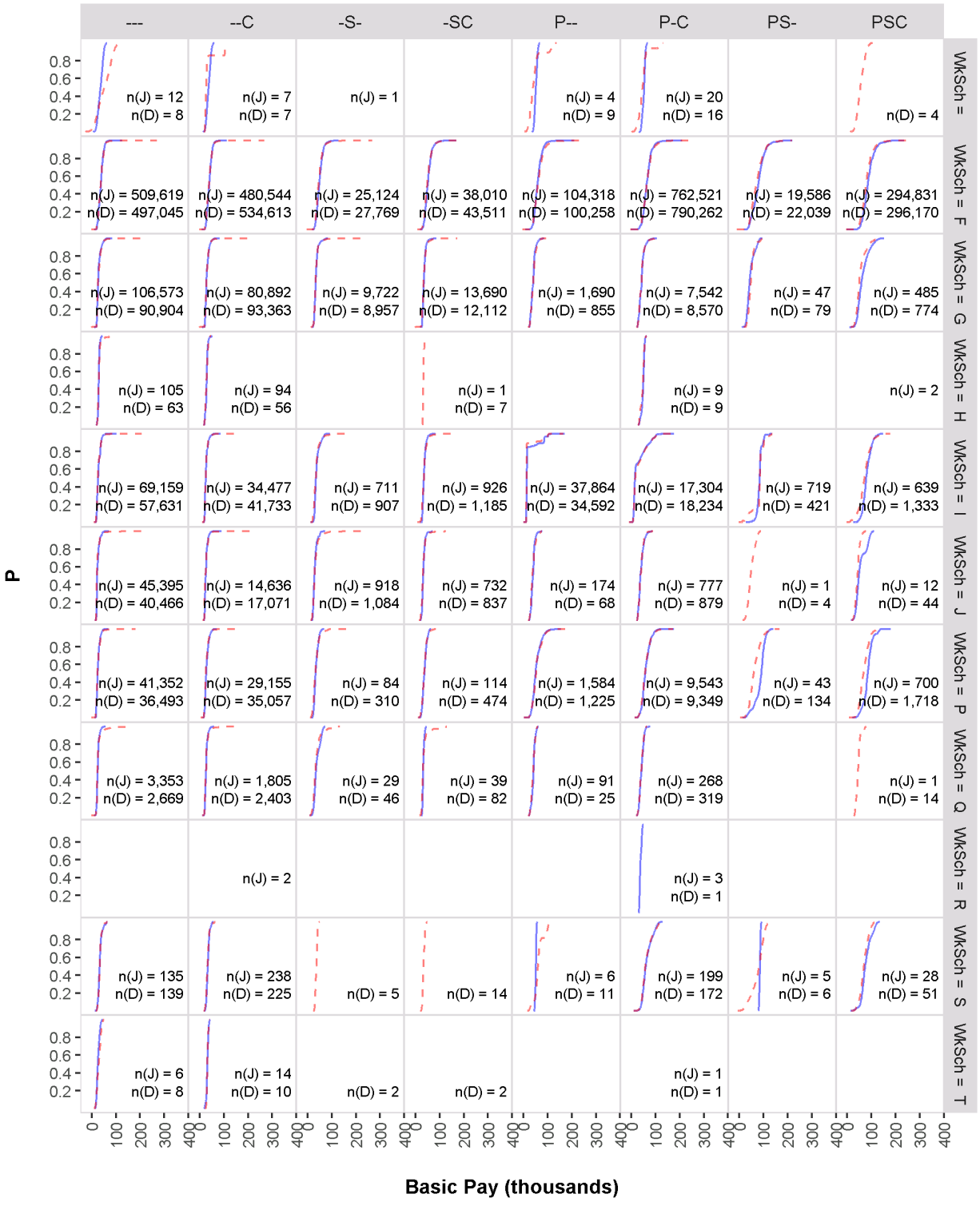


Figure 10: Basic pay distribution by joint professional, supervisory status, college education, and work schedule categories. Department of Agriculture (AG). Dashed line for synthetic data, solid line for authentic.

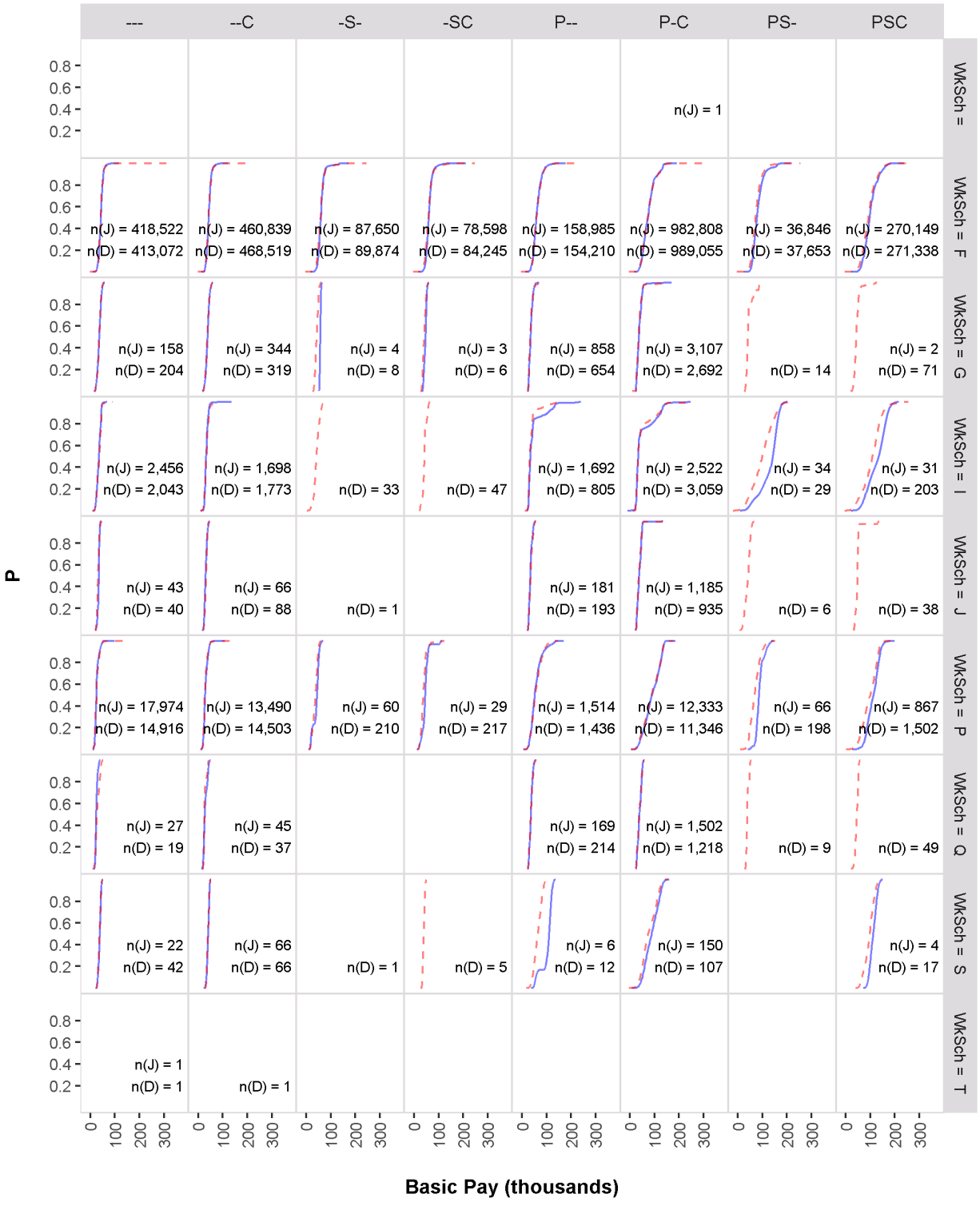


Figure 11: Basic pay distribution by joint professional, supervisory status, college education, and work schedule categories. Department of Justice (DJ). Dashed line for synthetic data, solid line for authentic.

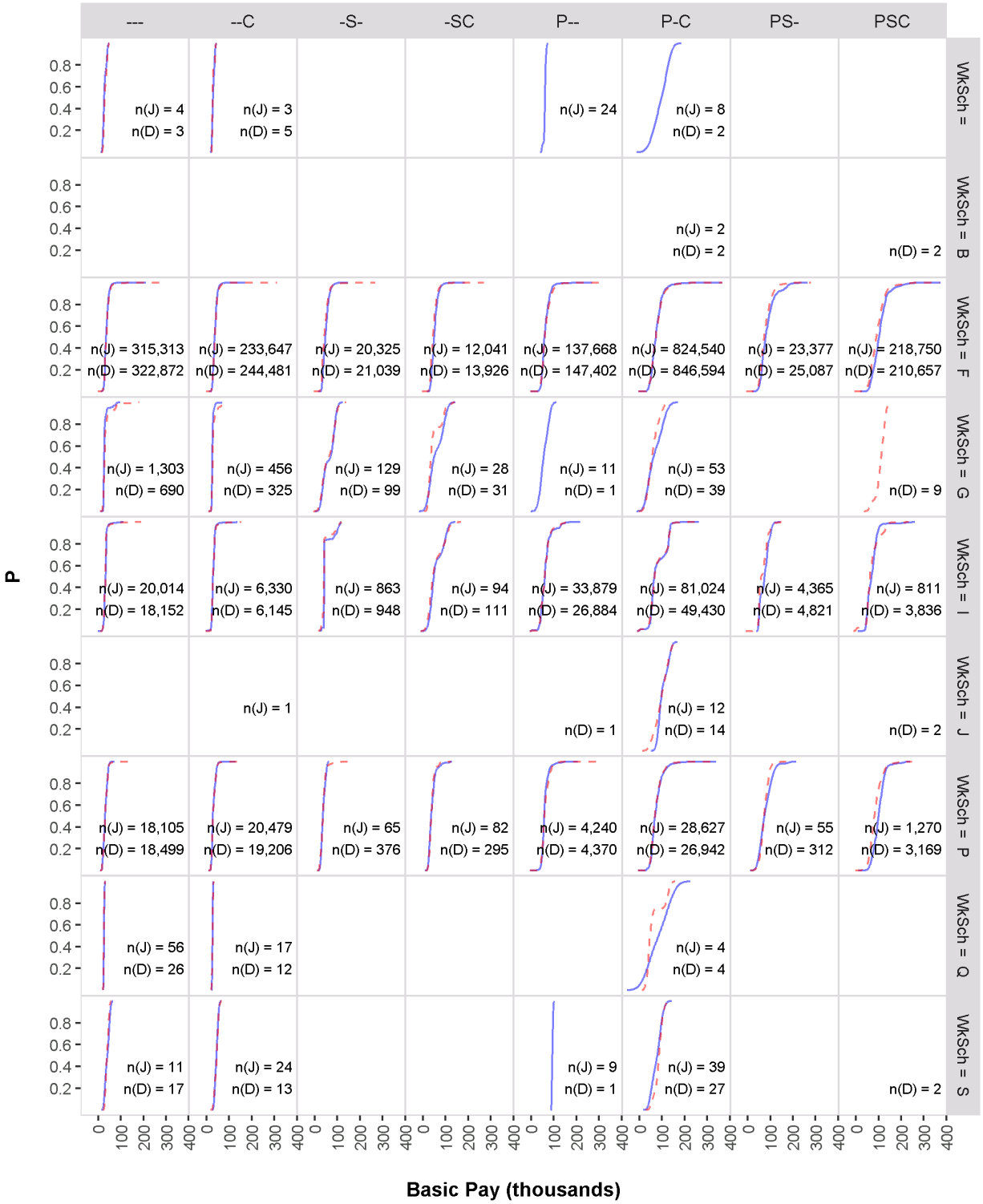


Figure 12: Basic pay distribution by joint professional, supervisory status, college education, and work schedule categories. Department of Health and Human Services (HE). Dashed line for synthetic data, solid line for authentic.

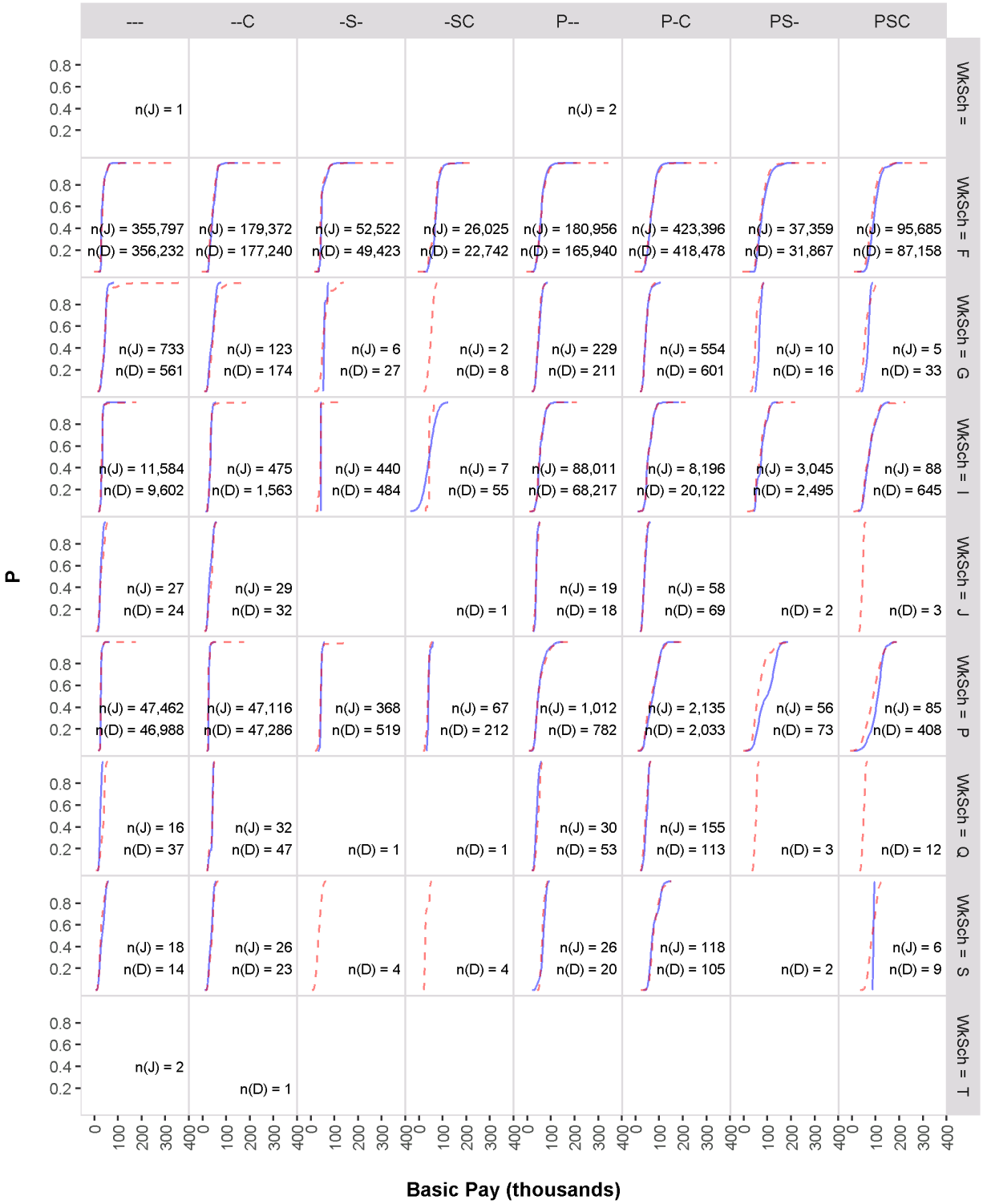


Figure 13: Basic pay distribution by joint professional, supervisory status, college education, and work schedule categories. Department of Homeland Security (HS). Dashed line for synthetic data, solid line for authentic.

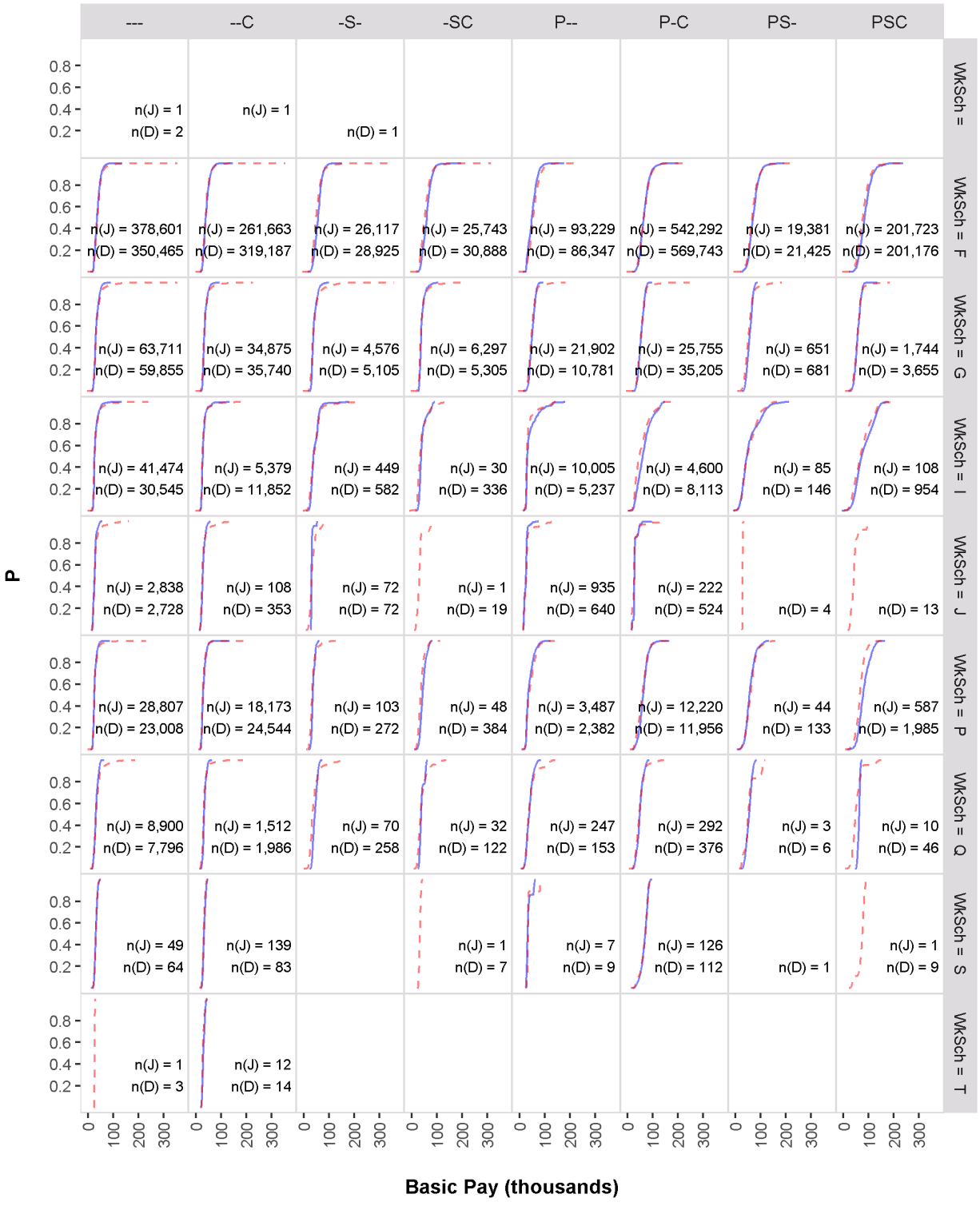


Figure 14: Basic pay distribution by joint professional, supervisory status, college education, and work schedule categories. Department of Interior (IN). Dashed line for synthetic data, solid line for authentic.

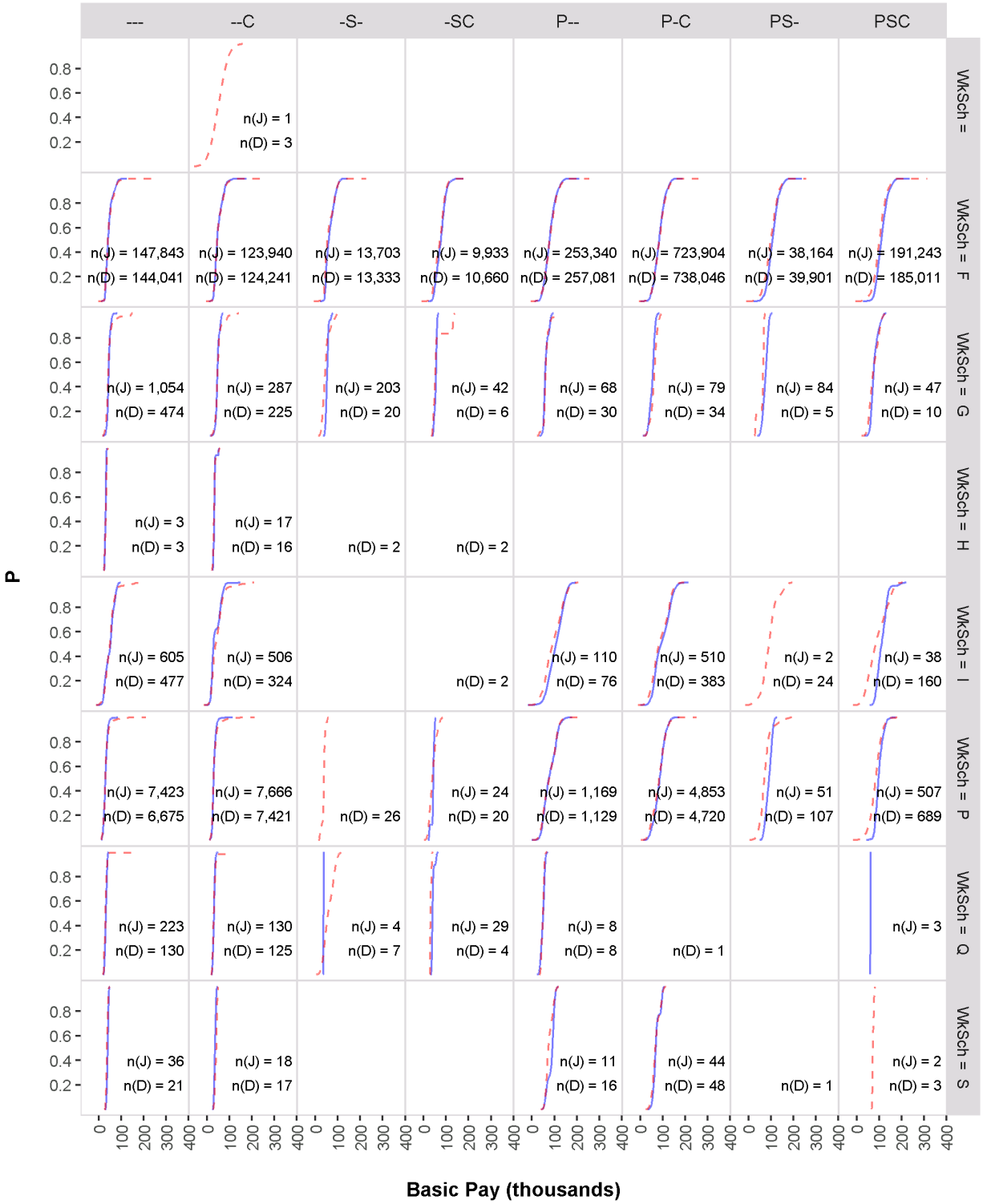


Figure 15: Basic pay distribution by joint professional, supervisory status, college education, and work schedule categories. Department of Transportation (TD). Dashed line for synthetic data, solid line for authentic.

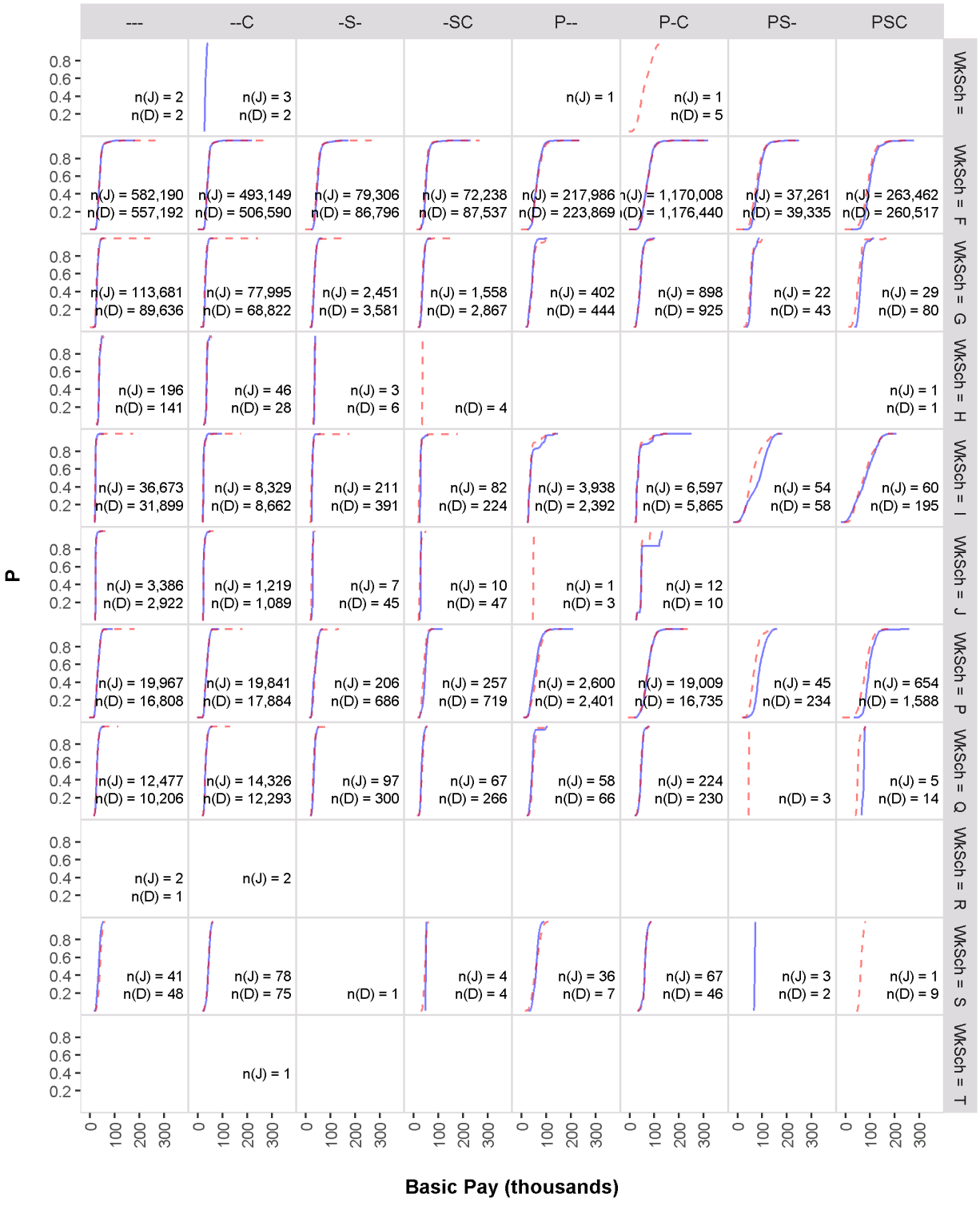


Figure 16: Basic pay distribution by joint professional, supervisory status, college education, and work schedule categories. Department of Treasury (TR). Dashed line for synthetic data, solid line for authentic.

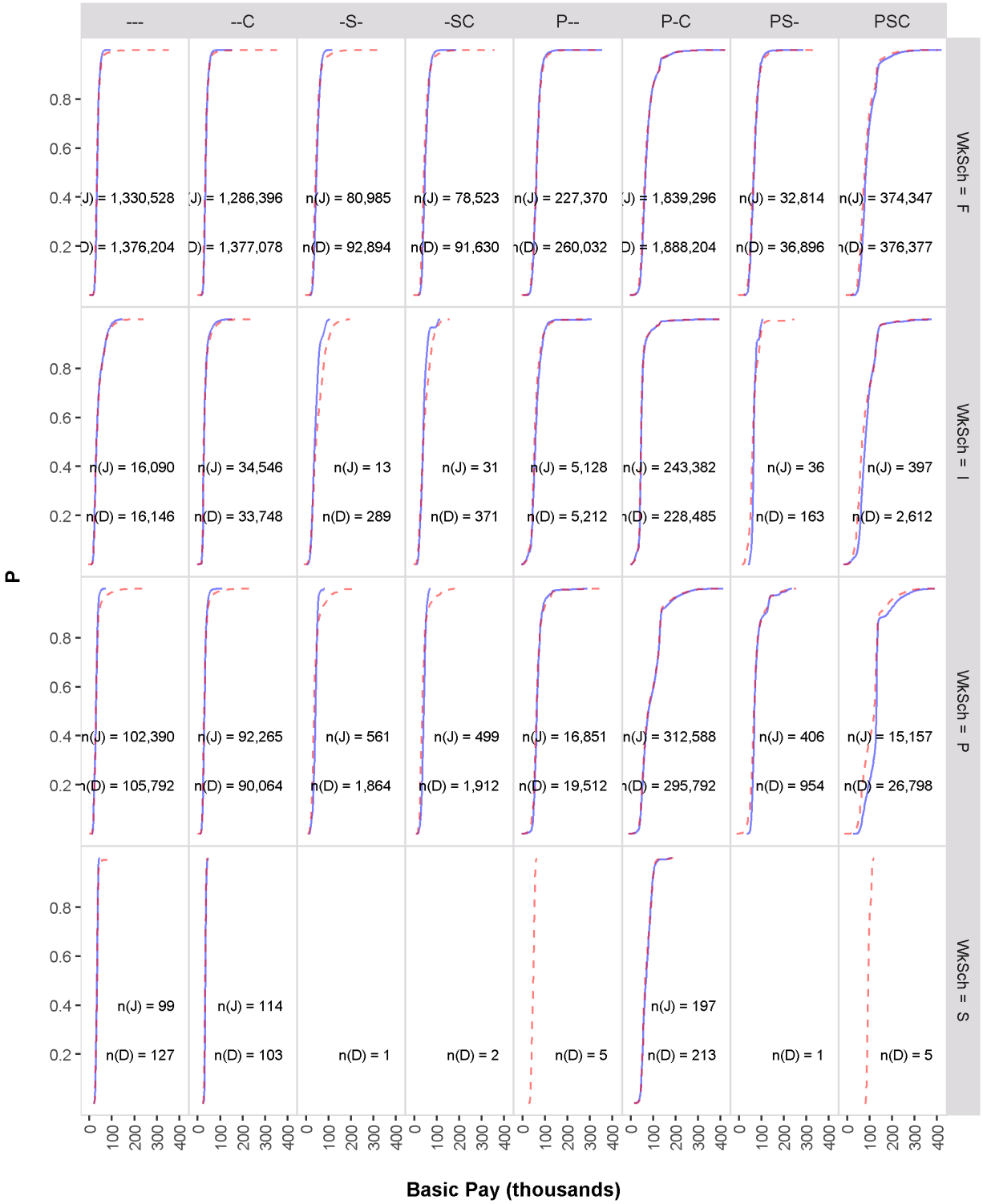
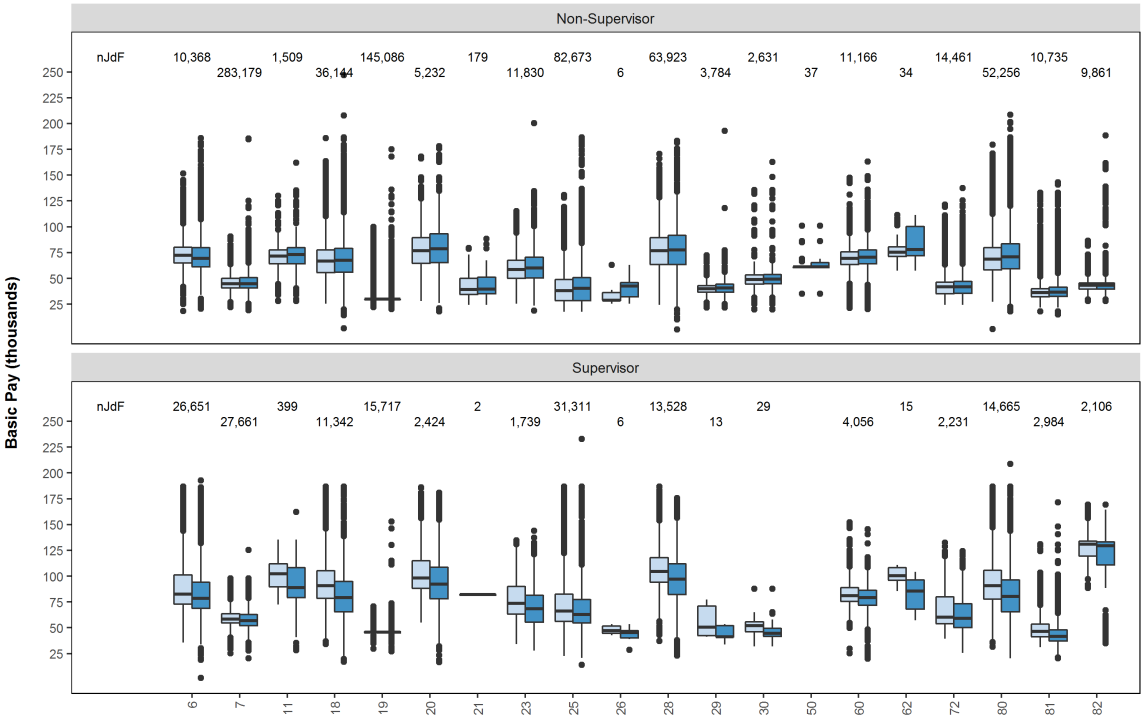


Figure 17: Basic pay distribution by joint professional, supervisory status, college education, and work schedule categories. Department of Veterans Affairs (VA). Dashed line for synthetic data, solid line for authentic.

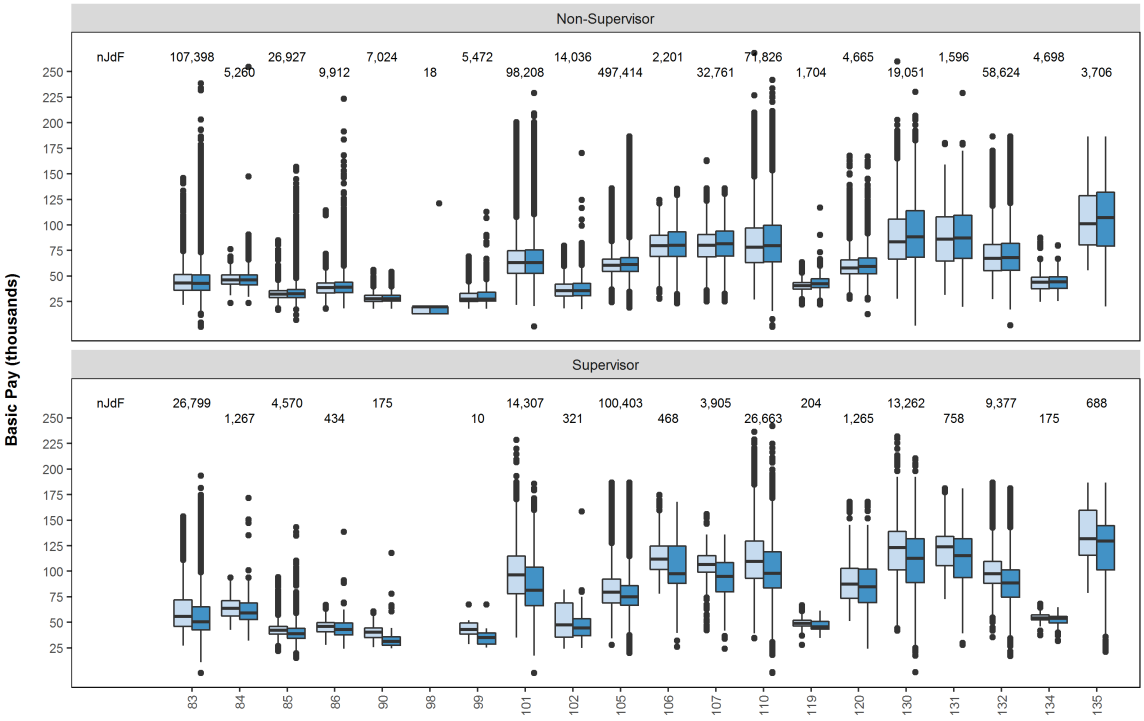
3.3 Distribution of Basic Pay by Occupation and Supervisory Status

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 18 through 20 show box plots for the first 120 occupations in order of occupation code. Trade, or blue collar, occupations begin at code 2500. Figures 21 and 22 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.

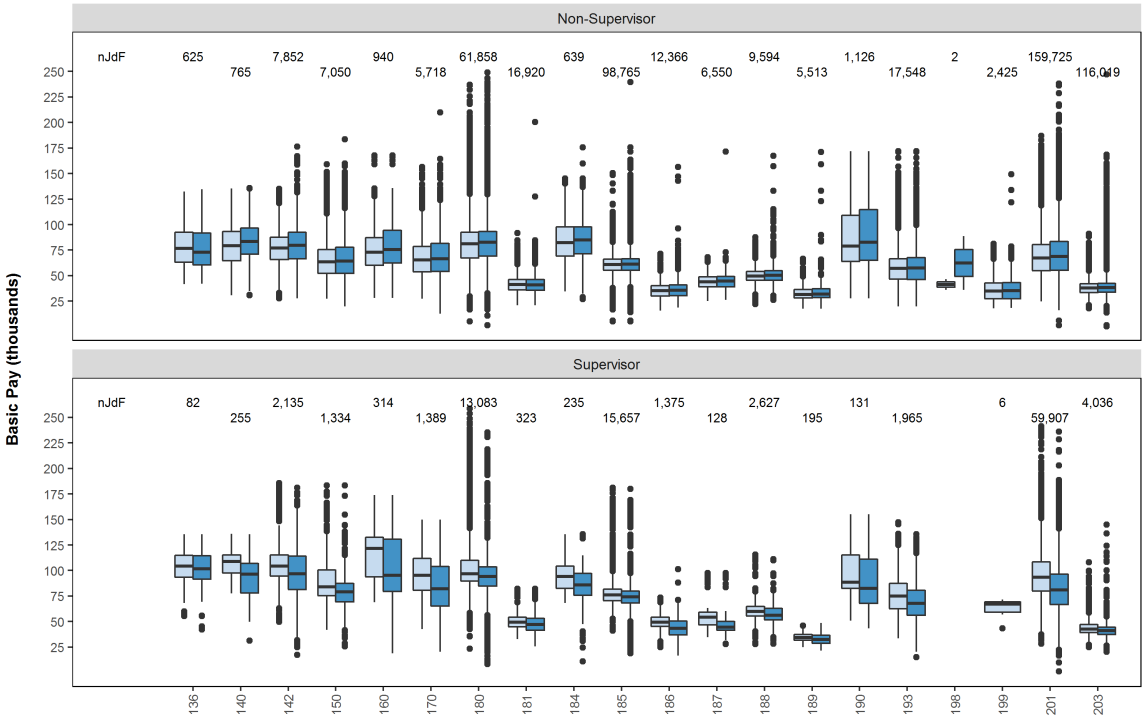


(a) Occupations 0006 through 0082

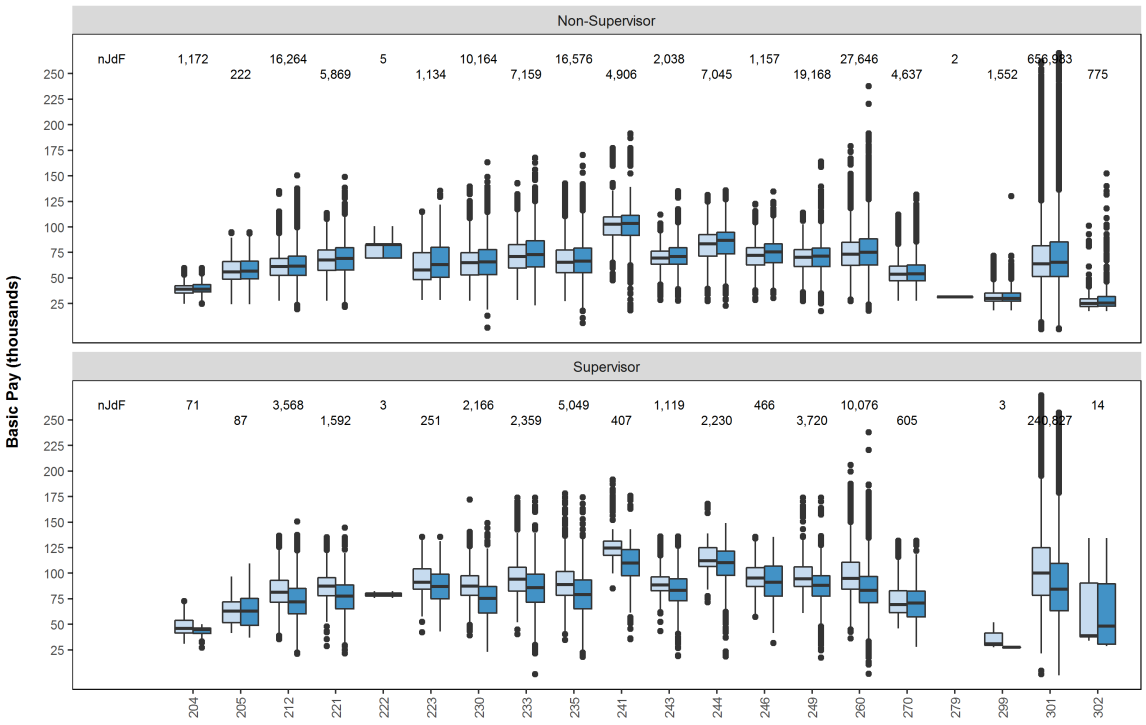


(b) Occupations 0083 through 0135

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

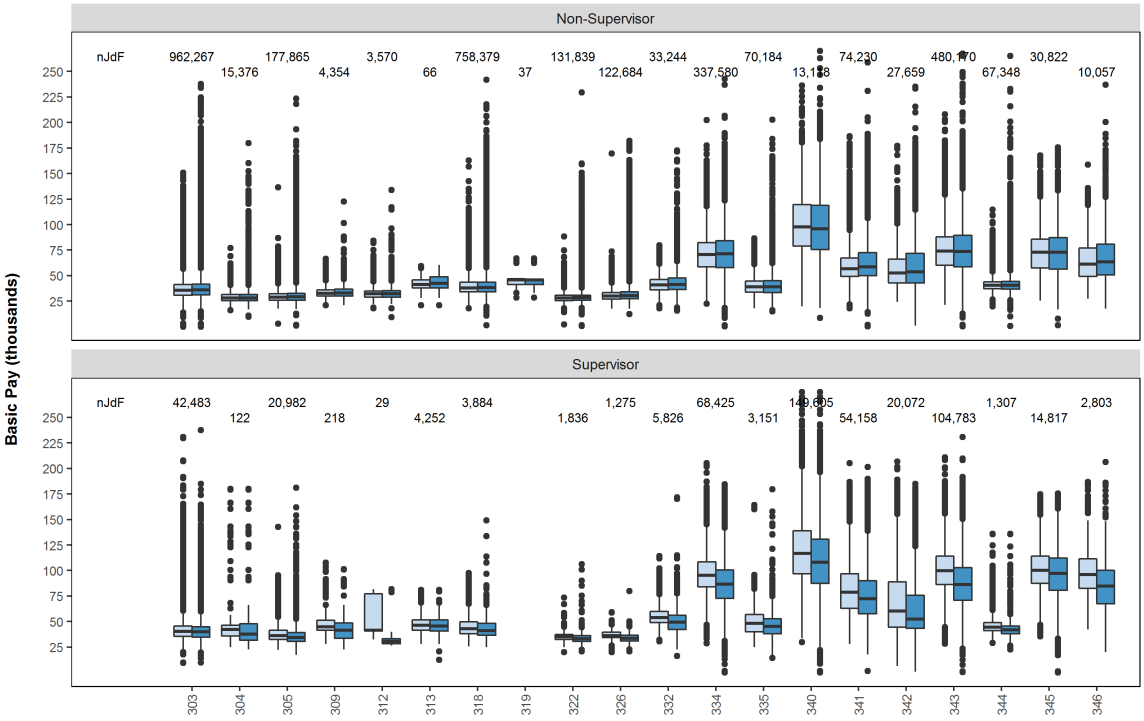


(a) Occupations 0136 through 0203

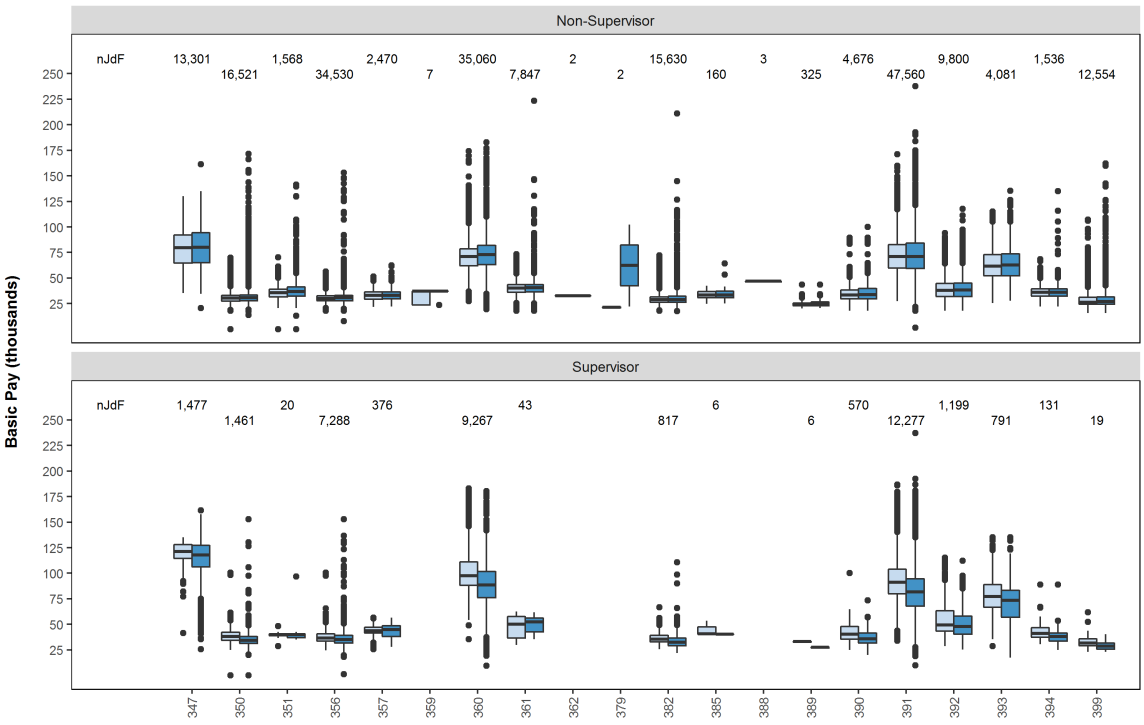


(b) Occupations 0204 through 0302

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

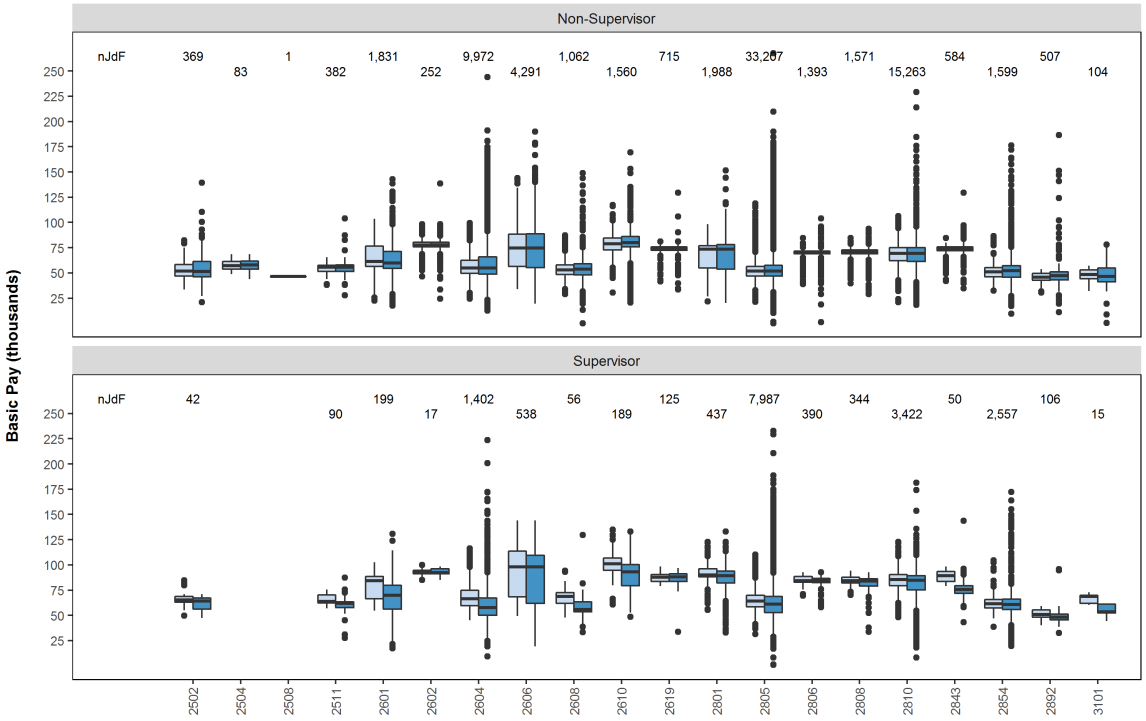


(a) Occupations 0303 through 0346

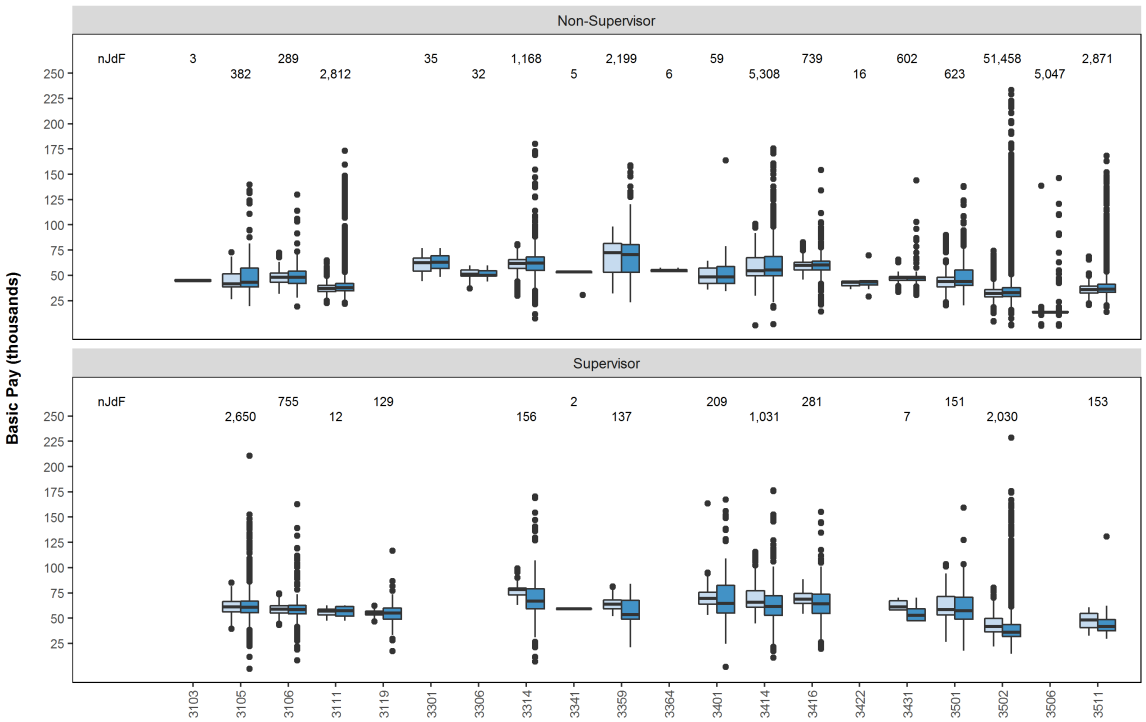


(b) Occupations 0347 through 0399

Figure 20: Basic pay distribution by occupation and supervisor status. All agencies combined. Authentic boxes on left, synthetic on right.

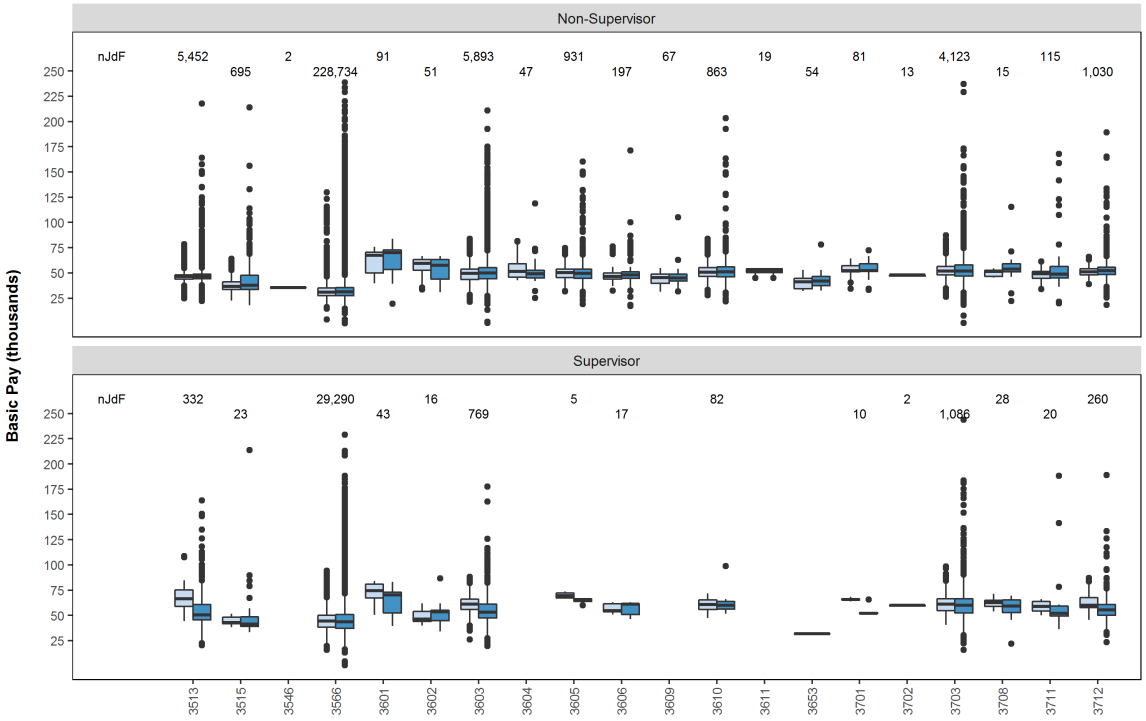


(a) Occupations 2502 through 3101 (trades)

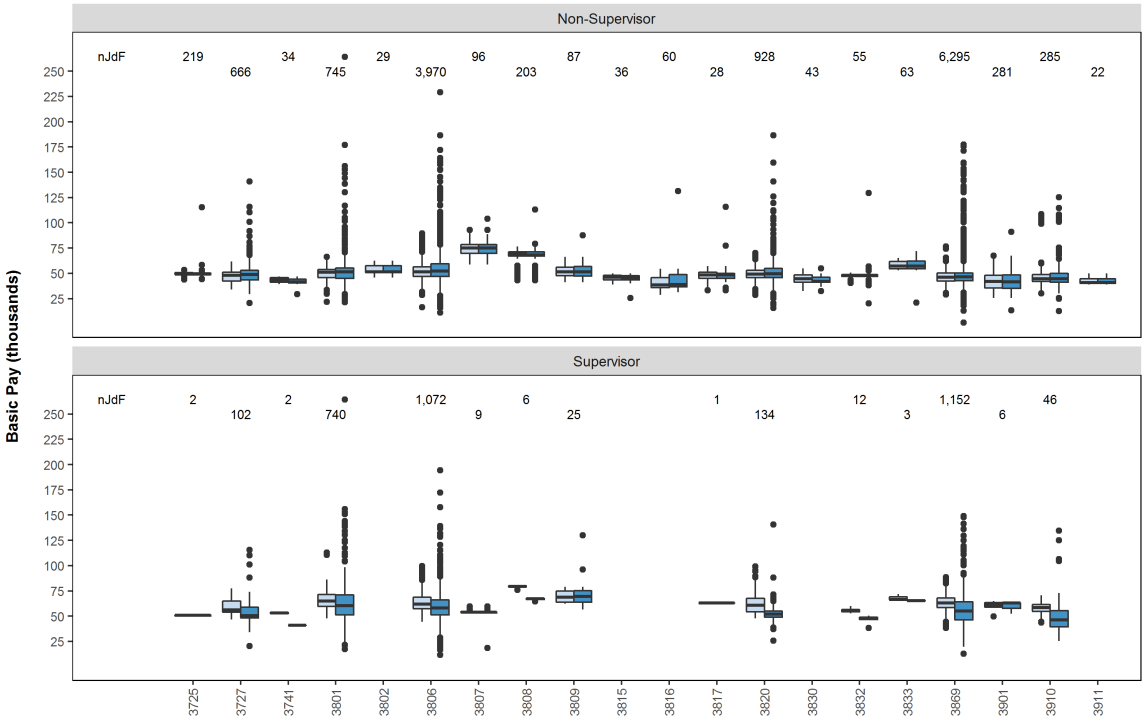


(b) Occupations 3103 through 3511 (trades)

Figure 21: 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 22: Basic pay distribution by occupation and supervisor status. All agencies combined. Authentic boxes on left, synthetic on right.

3.4 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 23 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.

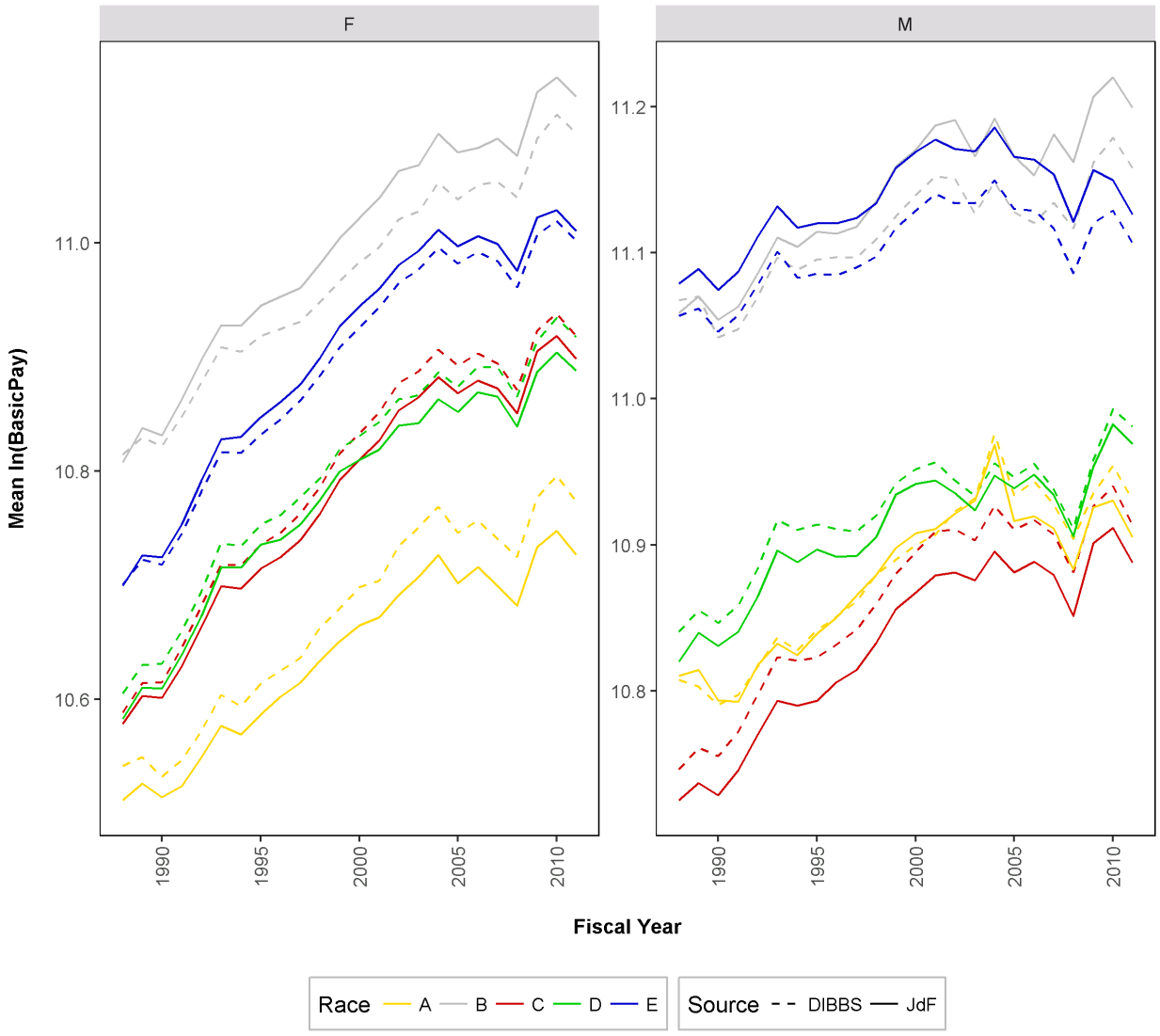


Figure 23: Mean basic pay by sex, race, and year. Female on left, male on right. Race codes: A = Native American, B = Asian, C = black, D = Hispanic, E = white. Dashed line for synthetic data, solid line for authentic.

4 Distribution of Gender

4.1 Gender Proportion by Race, Education, and Year

Accurate proportion observations by gender is critical in reproducing authentic results using models involving gender effects. Figure 24 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.

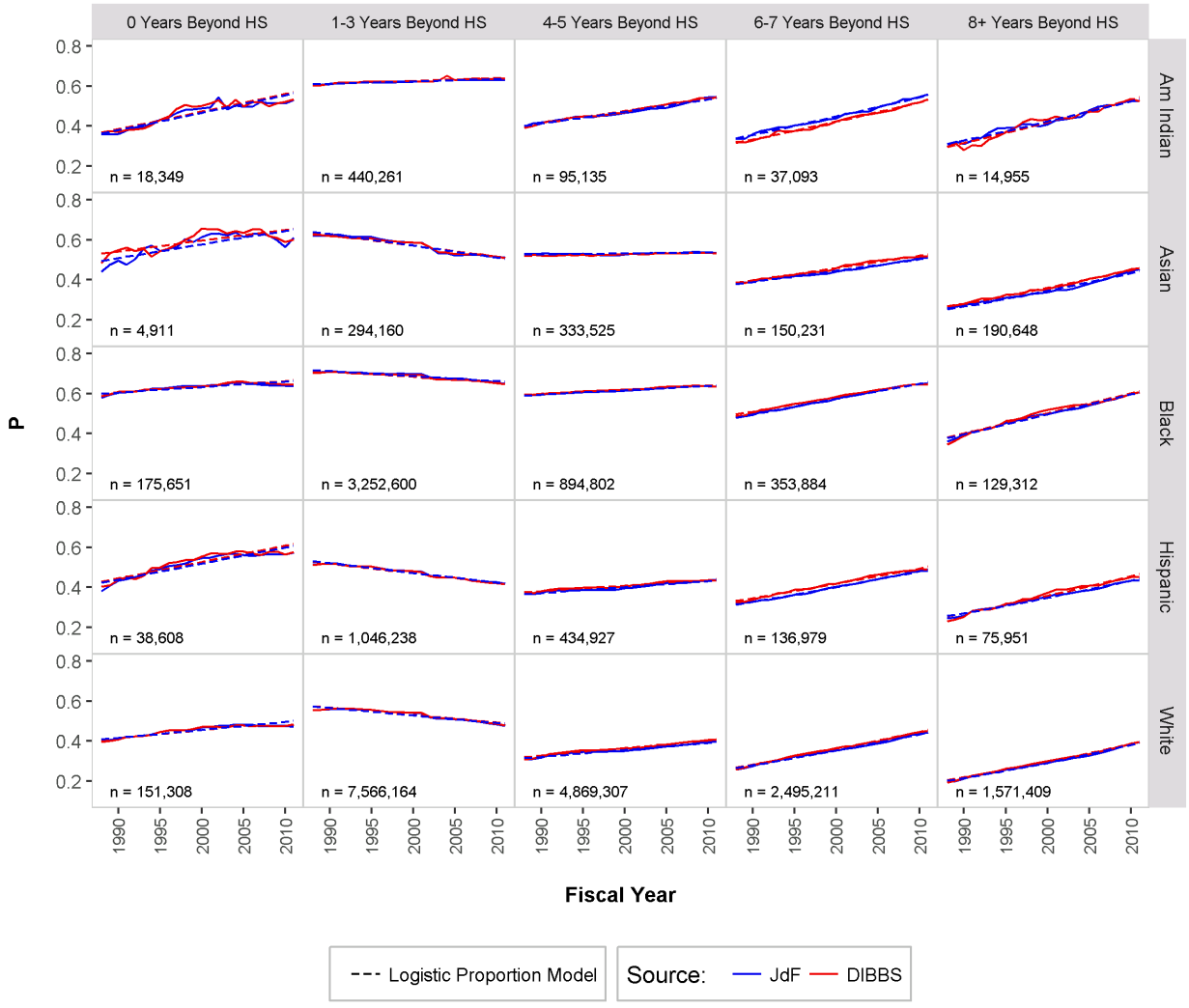


Figure 24: Proportion female observations by race, education, and year. Fitted lines are logistic regression estimates.

4.2 Gender Proportion by Race, Age, and Year

Figures 25 through 29, 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.

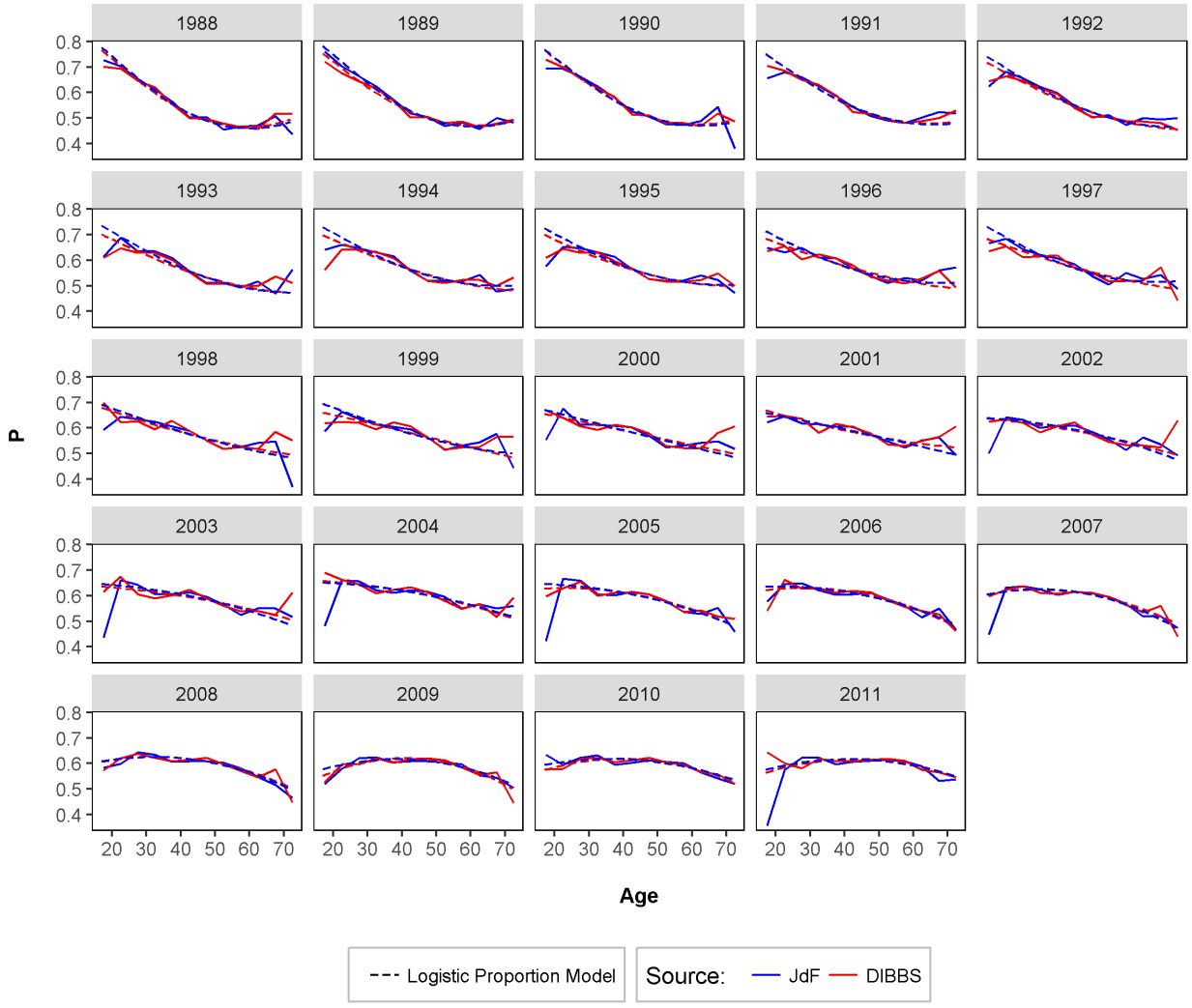


Figure 25: Proportion female observations by education and year. Race Native American. Fitted lines are logistic regression estimates.

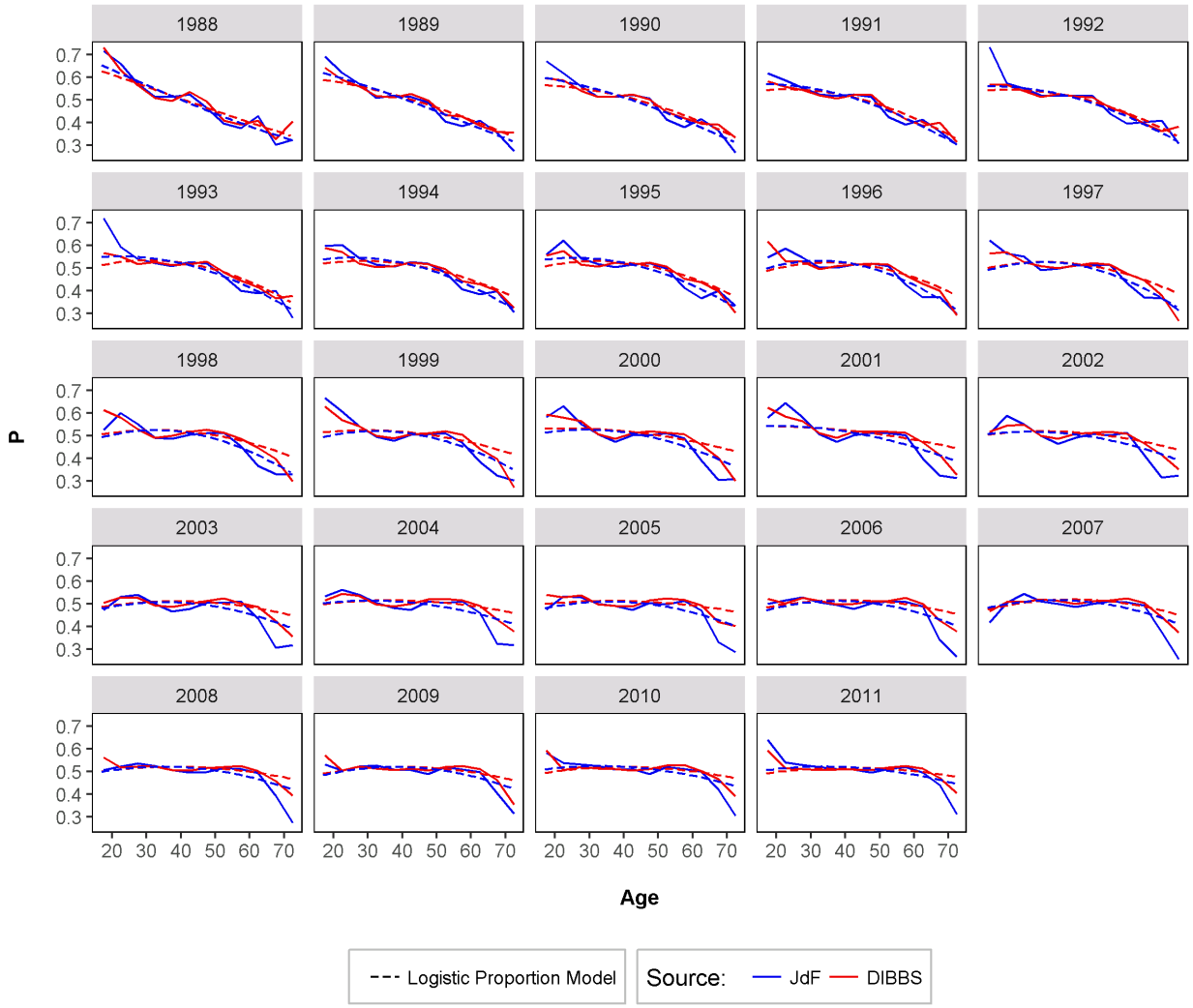


Figure 26: Proportion female observations by education and year. Race Asian. Fitted lines are logistic regression estimates.

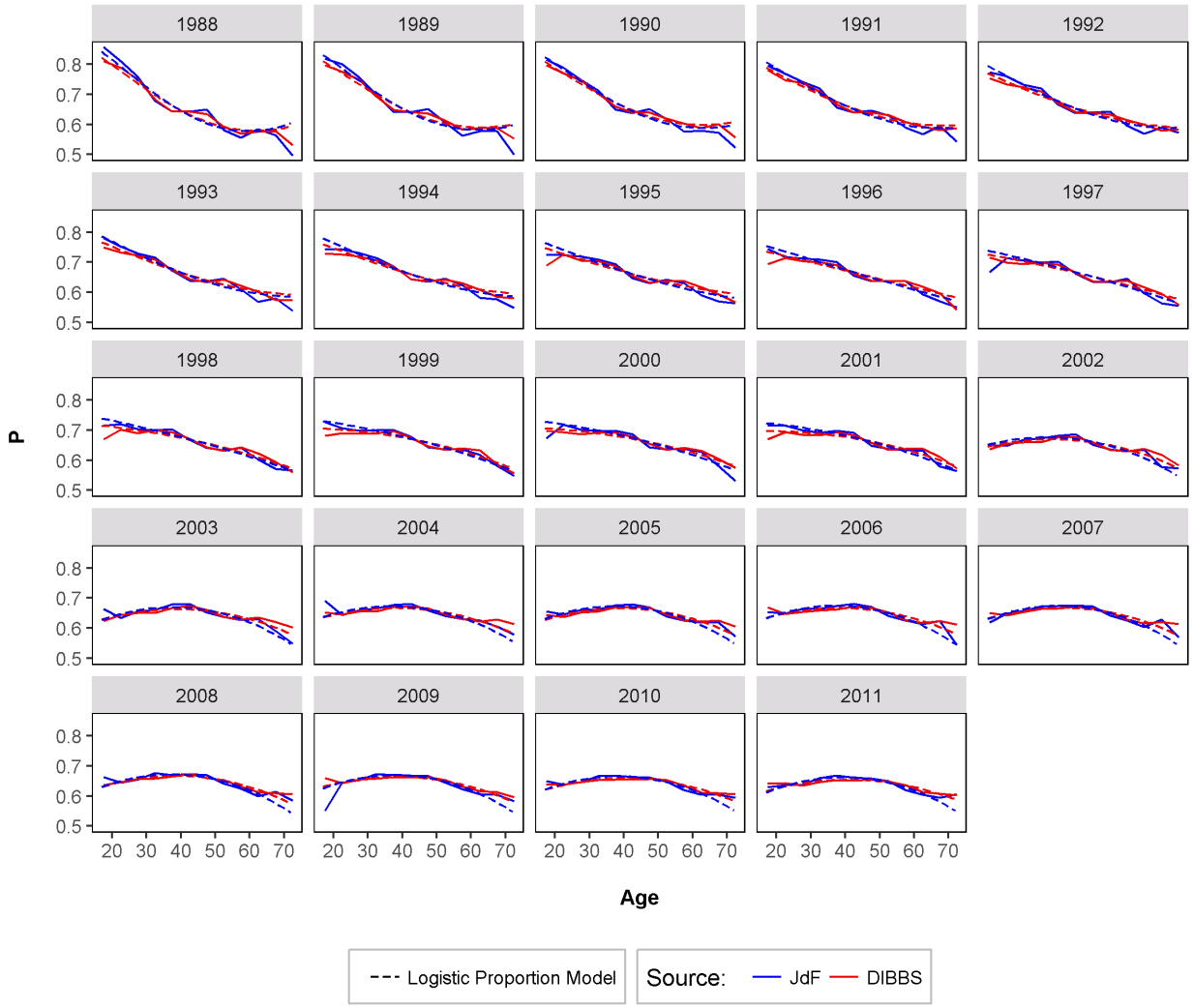


Figure 27: Proportion female observations by education and year. Race black. Fitted lines are logistic regression estimates.

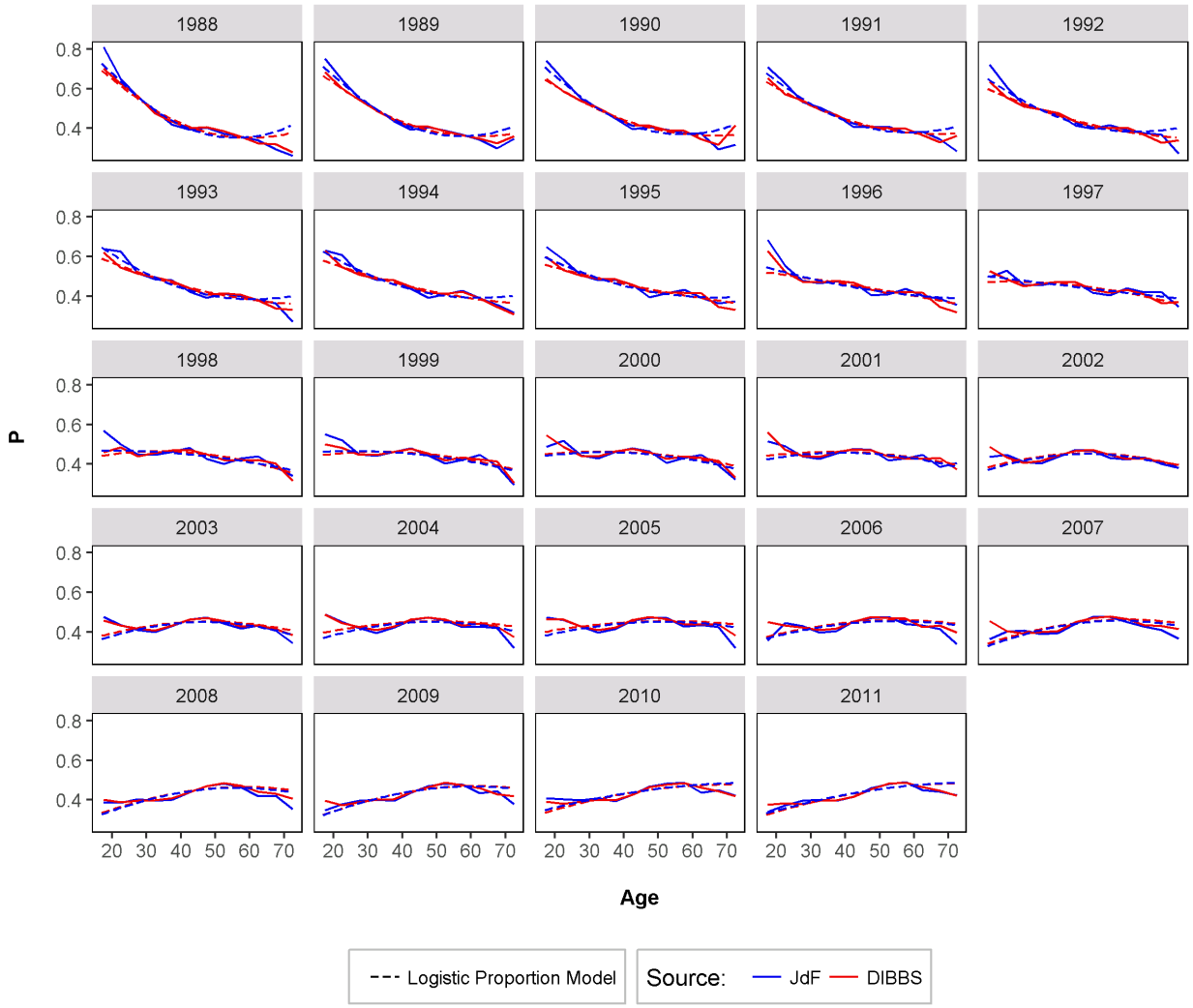


Figure 28: Proportion female observations by education and year. Race Hispanic. Fitted lines are logistic regression estimates.

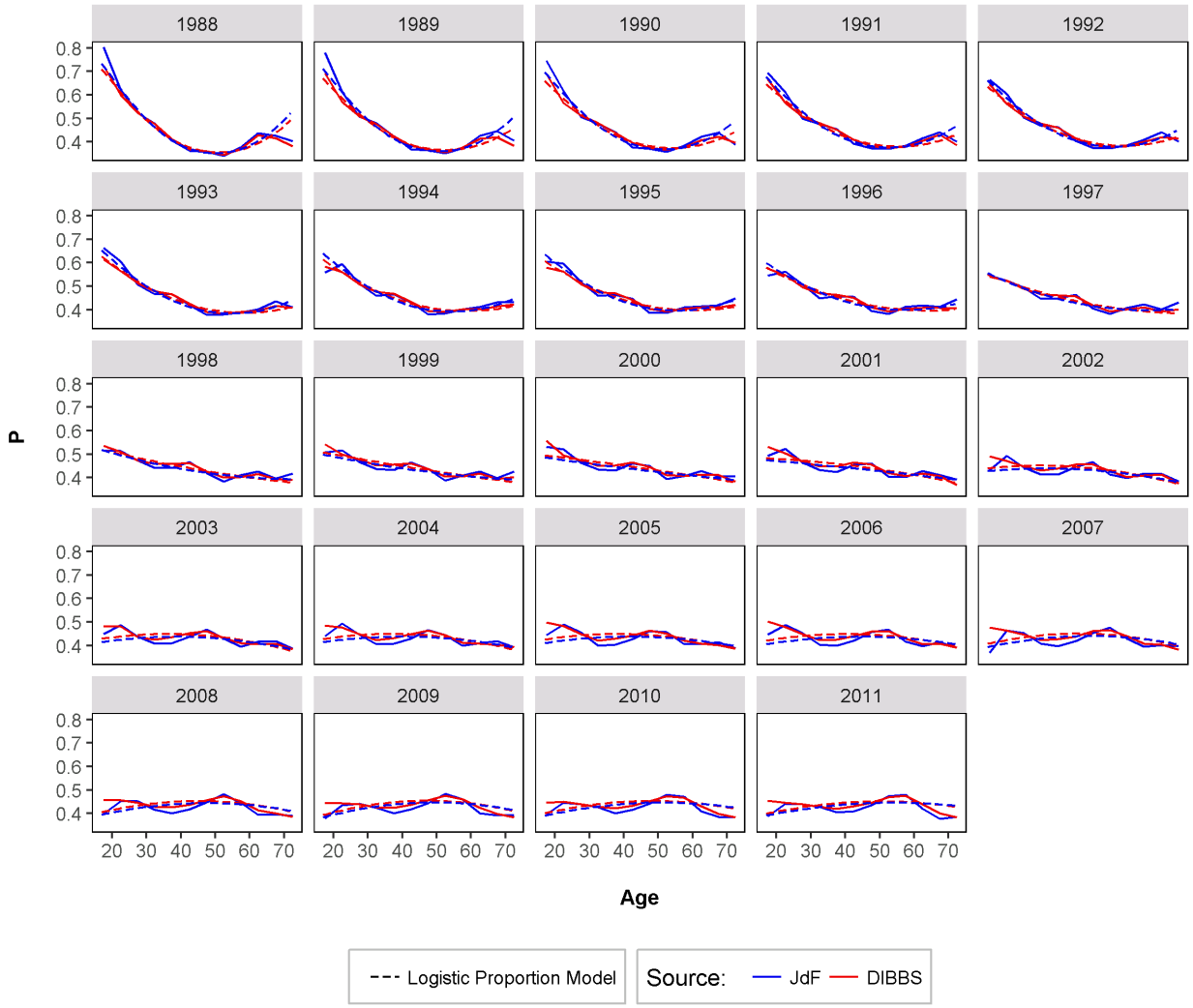
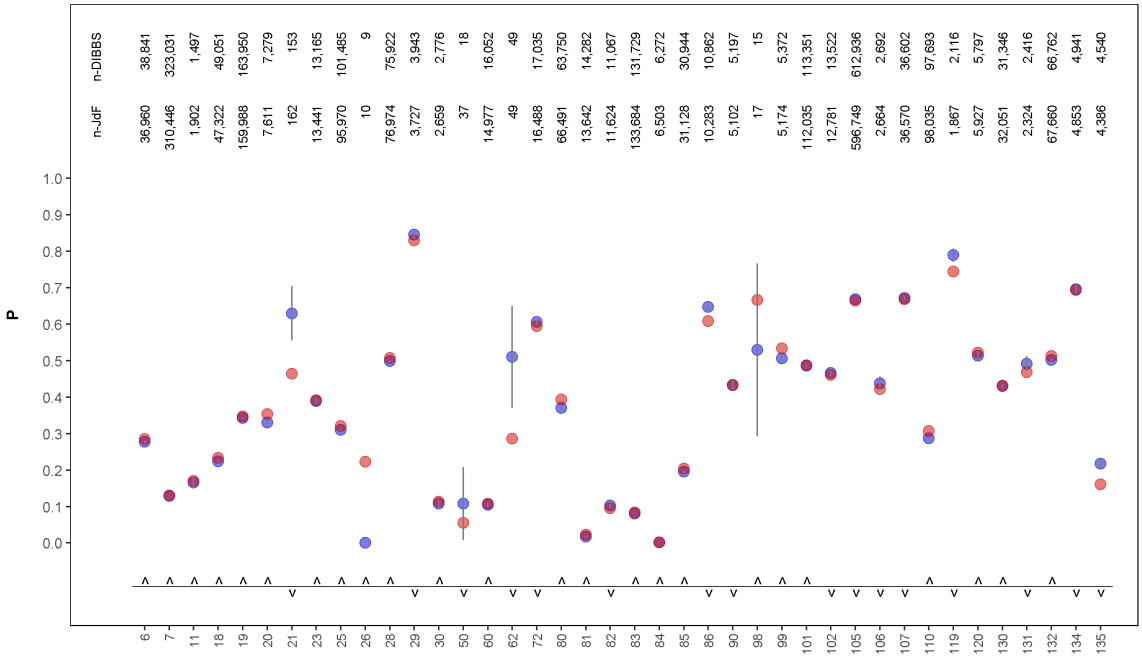


Figure 29: Proportion female observations by education and year. Race white. Fitted lines are logistic regression estimates.

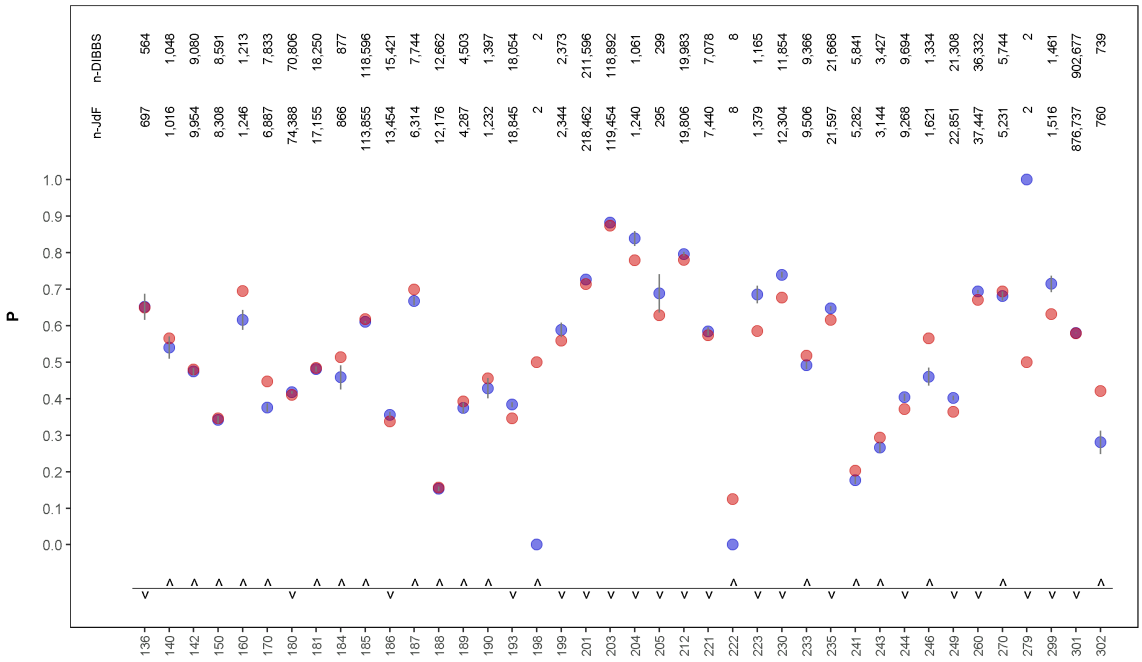
4.3 Gender Proportion by Occupation

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 30 and 31 compare proportion female for the first 120 occupation codes. Figures 32 and 33 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.

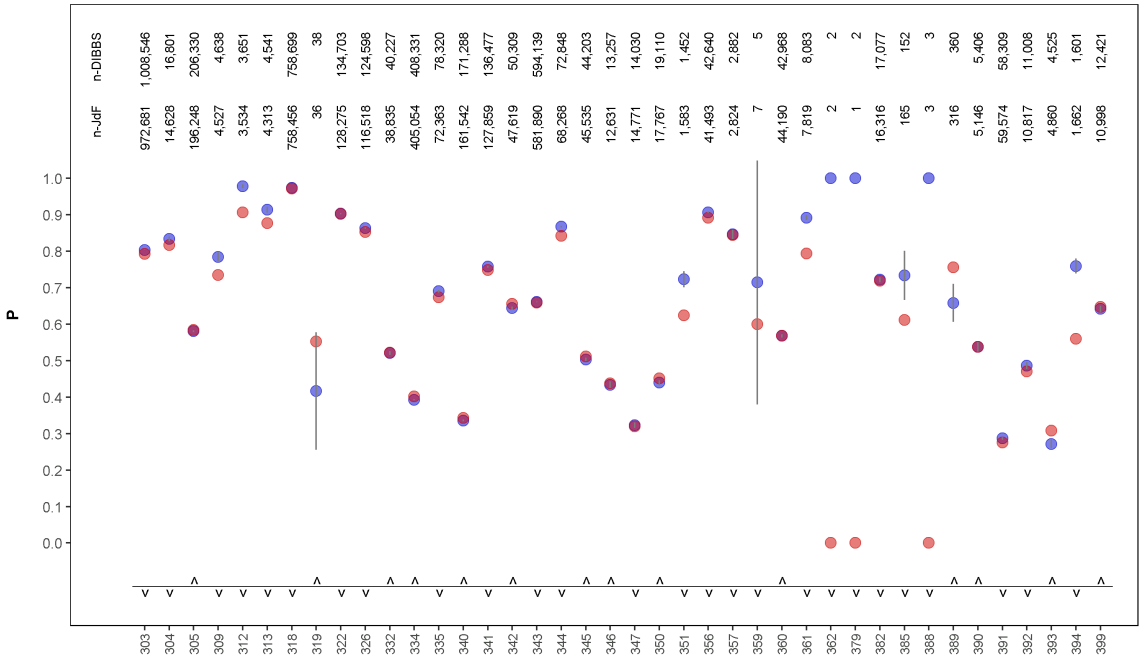


(a) Occupations 0006 through 0135

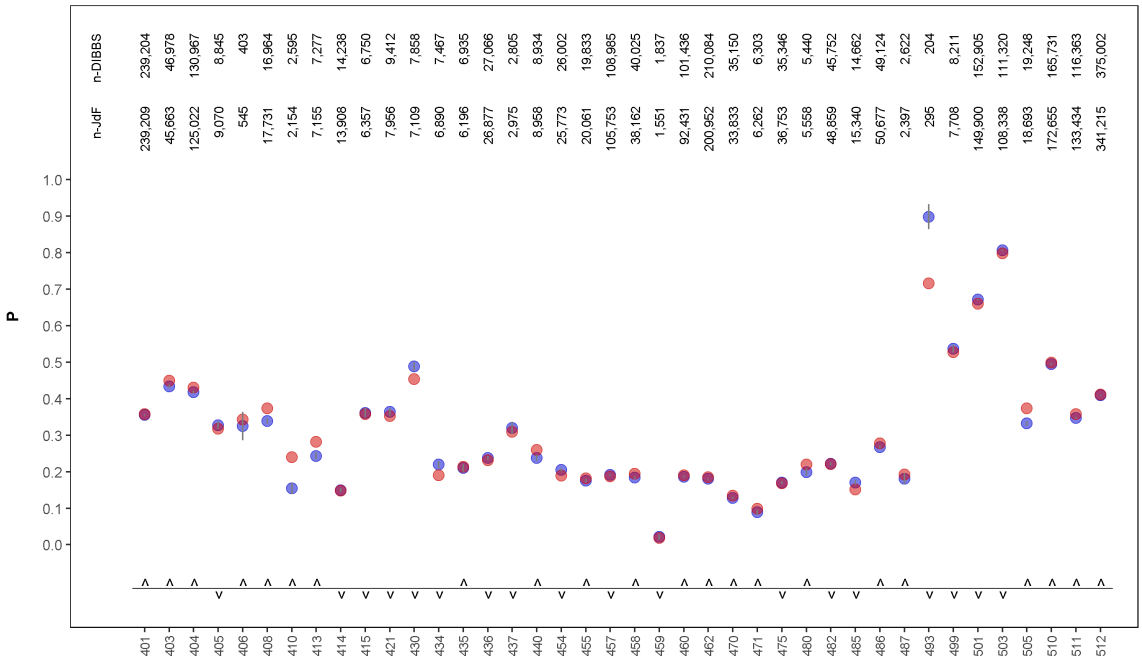


(b) Occupations 0136 through 0302

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

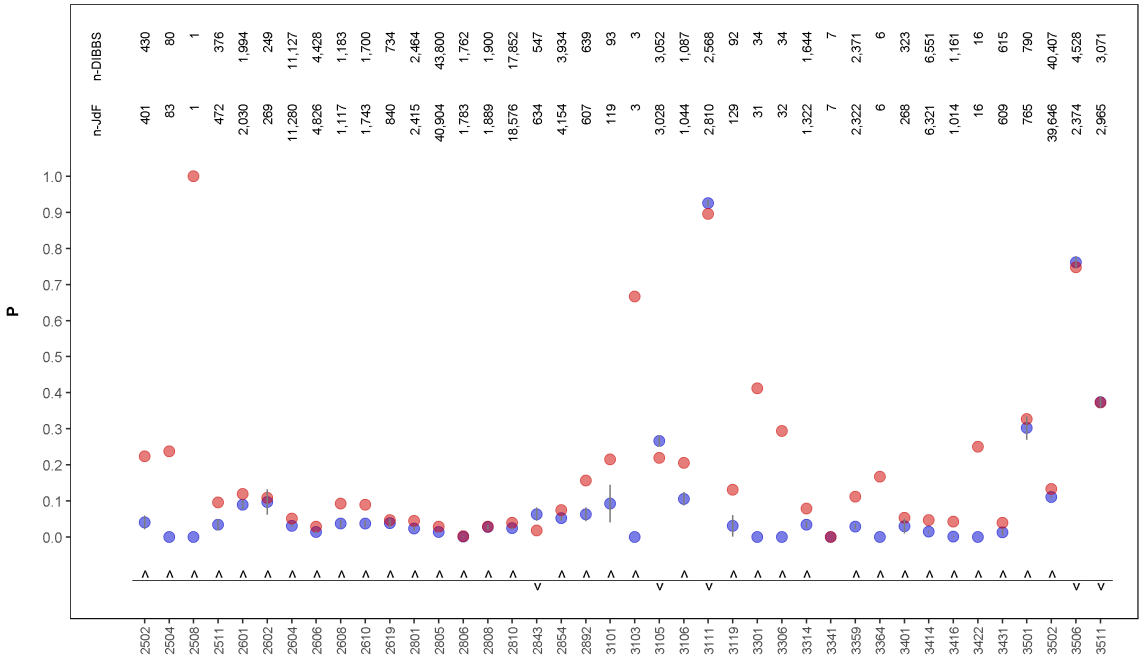


(a) Occupations 0303 through 0399

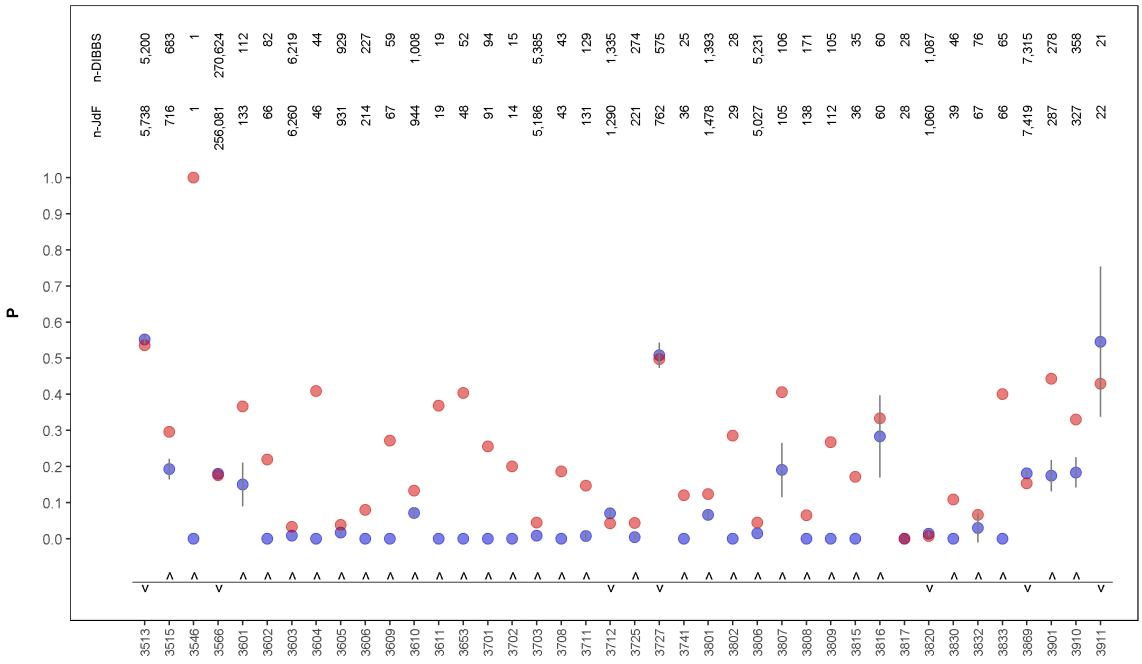


(b) Occupations 0401 through 0512

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

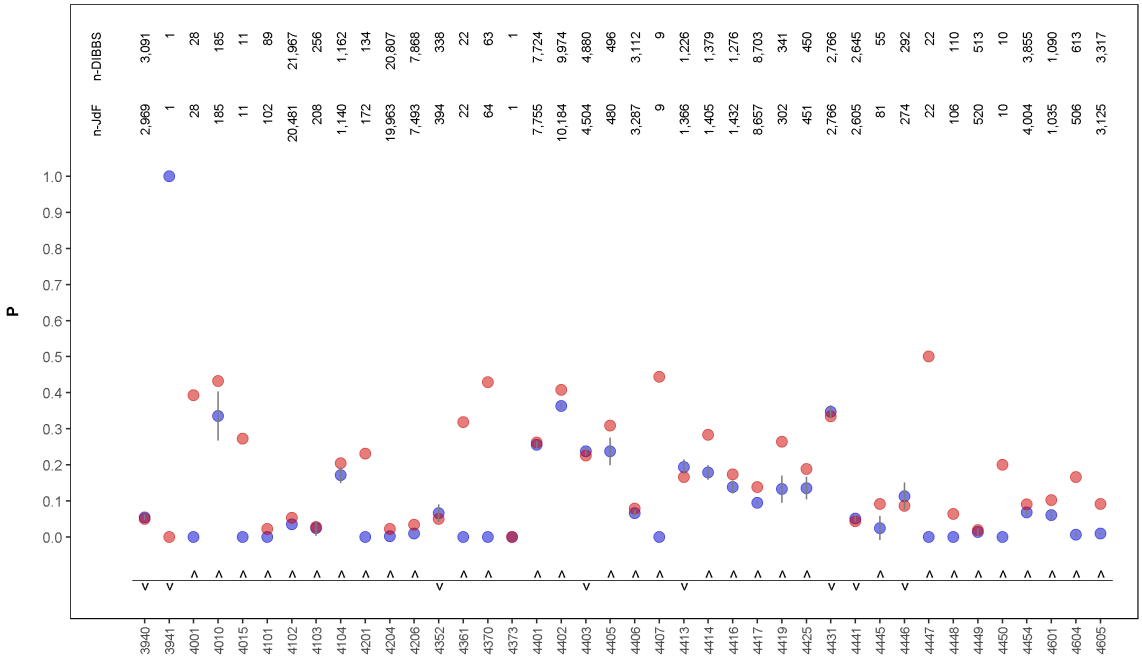


(a) Occupations 2502 through 3511 (trades)

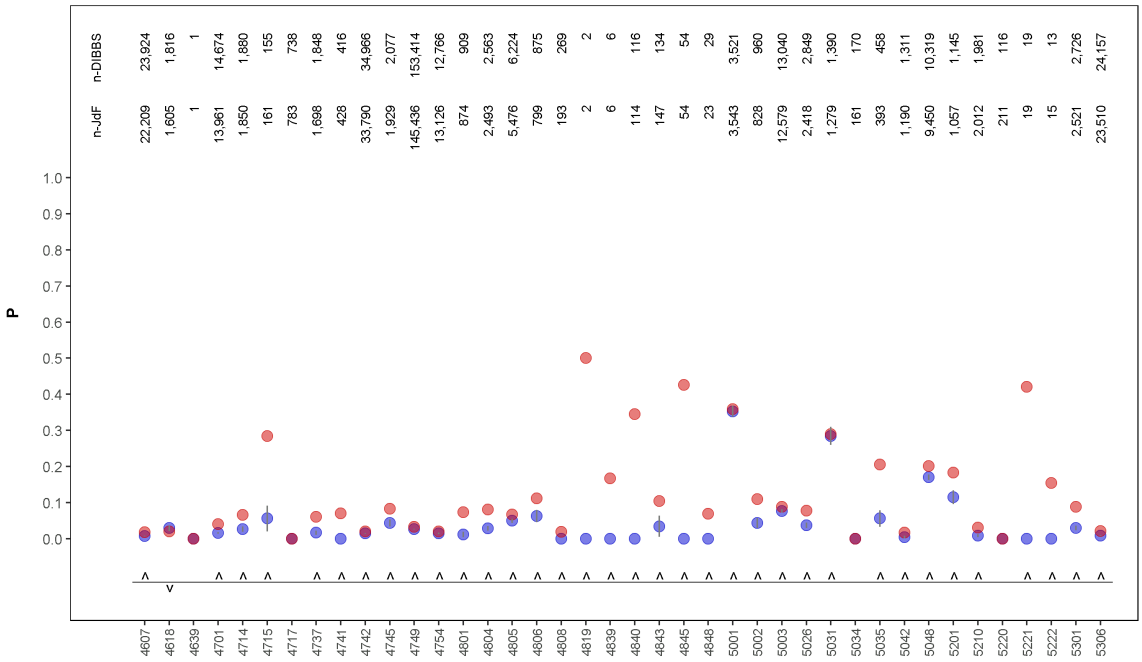


(b) Occupations 3513 through 3911 (trades)

Figure 32: 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 33: Proportion female observations by occupation. All agencies combined. One synthetic and one authentic point per occupation.

4.4 Occupation Gender Proportion Kernel Distribution

Figure 34 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.

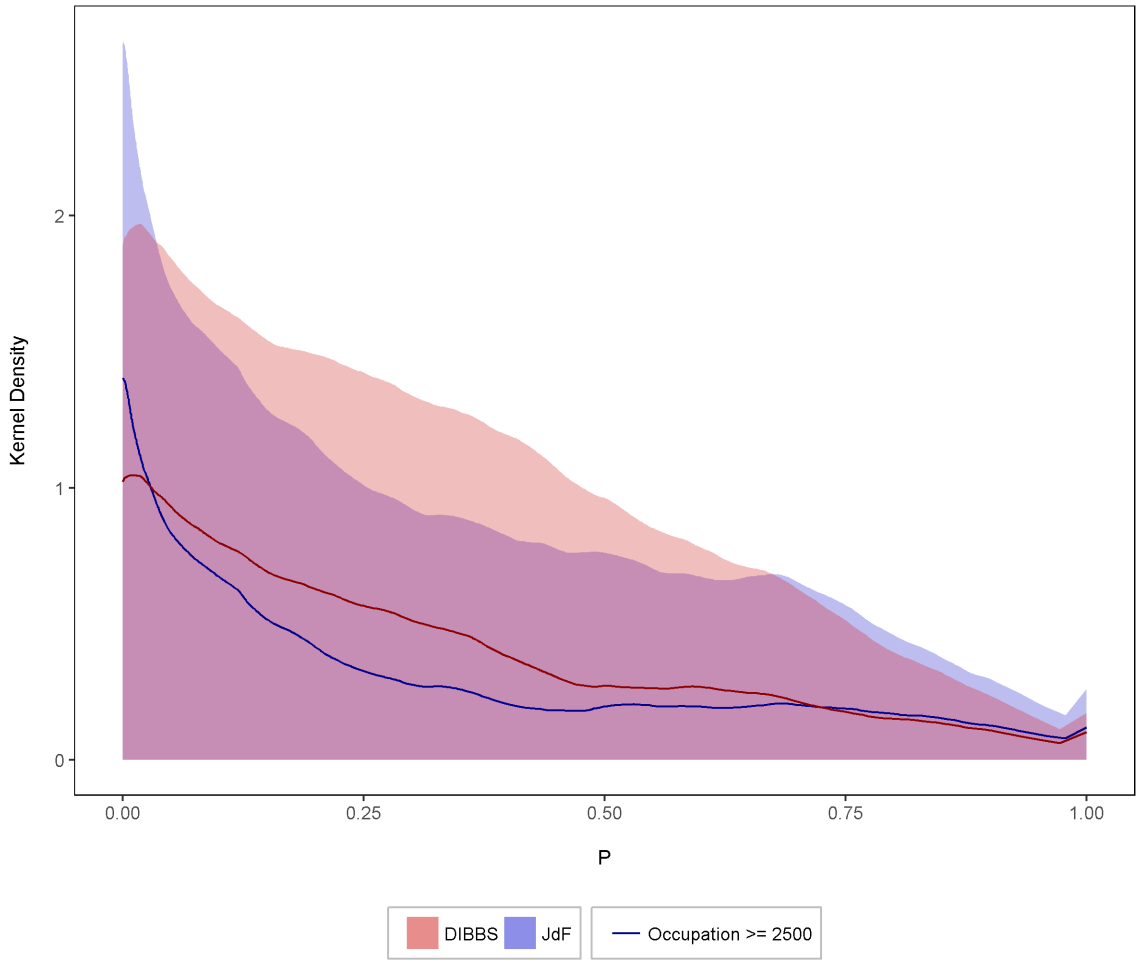


Figure 34: Occupation proportion female kernel density. Synthetic and authentic distributions superimposed. Trade occupations slightly over-represented near zero and above 0.75 in authentic data. Slight discrepancy in synthetic proportions at extremes. Compensated for near central proportions.

4.5 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 35 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.

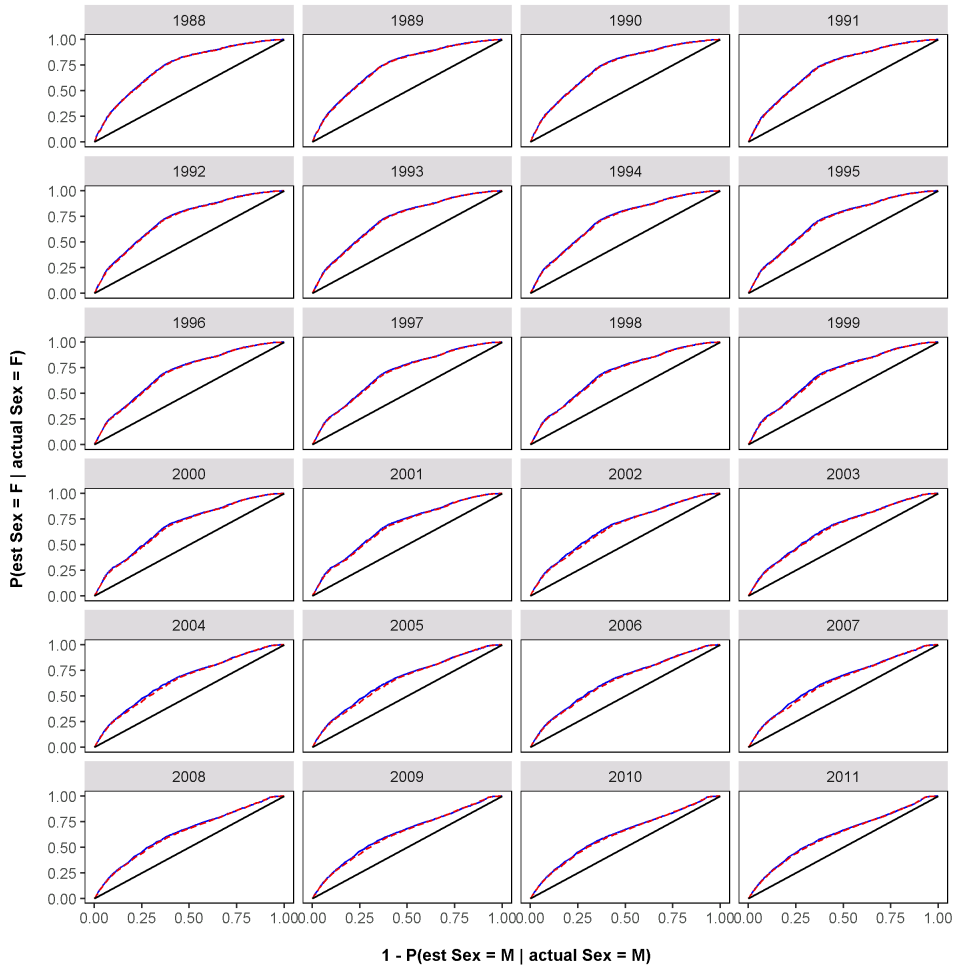


Figure 35: Proportion female race, age, education, occupation classifier ROC curves. One curve per data set per by fiscal. Agreement in classifier accuracy indicated by overlapping curves. Pattern of decreased accuracy as years progress captured in both data sets.

5 Gender Pay Differential Fixed Effects Quantile Regression Model

Disparity in pay by gender is an important and common topic in human capital research. Ordinary least squares regression can estimate the effect of gender on expected values of difference in pay, but we are also interested in 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 that estimates the effect of gender on pay quantiles for a given year, controlling for race, age, education, agency, and occupation is

$$\hat{\boldsymbol{y}} = \hat{\boldsymbol{\beta}}_0 + \hat{\boldsymbol{\beta}}_s sex + \hat{\boldsymbol{\beta}}_r race + \hat{\boldsymbol{\beta}}_{ag} age + \hat{\boldsymbol{\beta}}_{ag^2} age^2 + \hat{\boldsymbol{\beta}}_{ed} education + \hat{\boldsymbol{\beta}}_{agcy} agency + \hat{\boldsymbol{\beta}}_{occ} occupation \qquad (1)$$

where $\hat{\boldsymbol{y}}$ is a particular quantile of log(basic pay). Figure 36 plots gender effects ($\hat{\boldsymbol{\beta}}_s$) from model (1) fit to annual subsets of observations for quantiles 0.1, 0.5, and 0.9.

Observations: Pay disparity is greater in all years for higher quantiles. Although some systematic separation appears between data sets, trends in the synthetic $\hat{\boldsymbol{\beta}}_s$ estimates convey the important finding of gradual decreased effect of gender for all three quantiles, to near parity at the end of the study period.

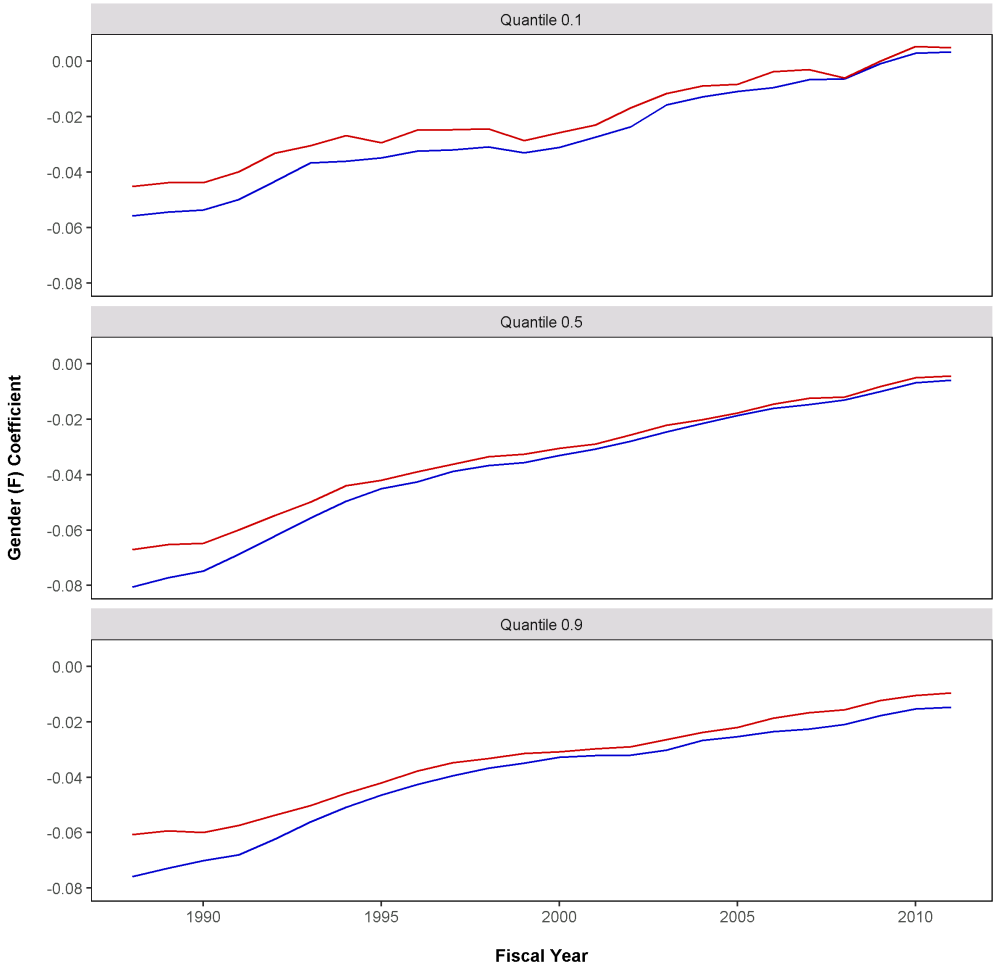


Figure 36: Pay disparity gender effect quantile estimates. Change over time. Upper line represents synthetic data, lower line authentic data.

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

This section uses results from a study of trends in pay grade 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."

6.1 Federal Wage Bill Decomposition

Figure 37 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.

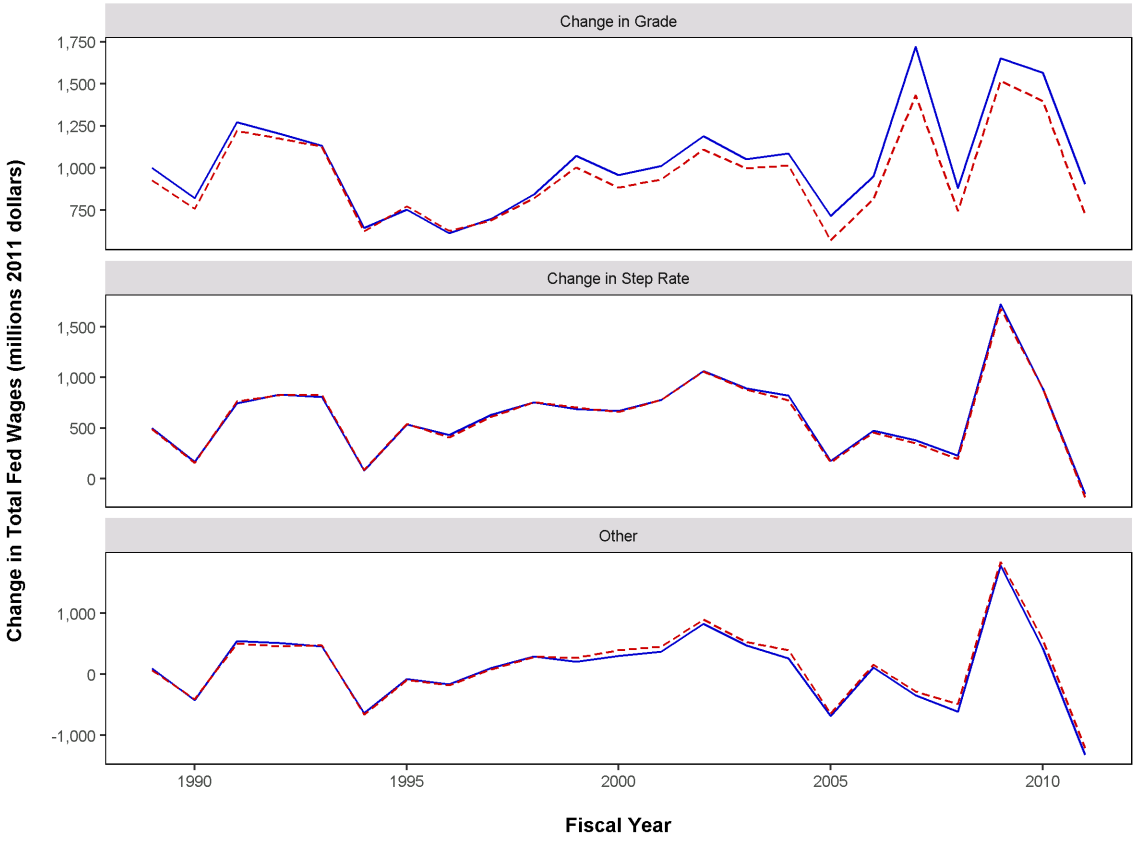


Figure 37: Change in U.S. federal government total wage bill. Millions of 2011 dollars by change in grade, change in step rate, and other changes. Fiscal years 1988 through 2011.

6.2 Change in GS Grade Distribution 2011 vs. 1988

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 38 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.

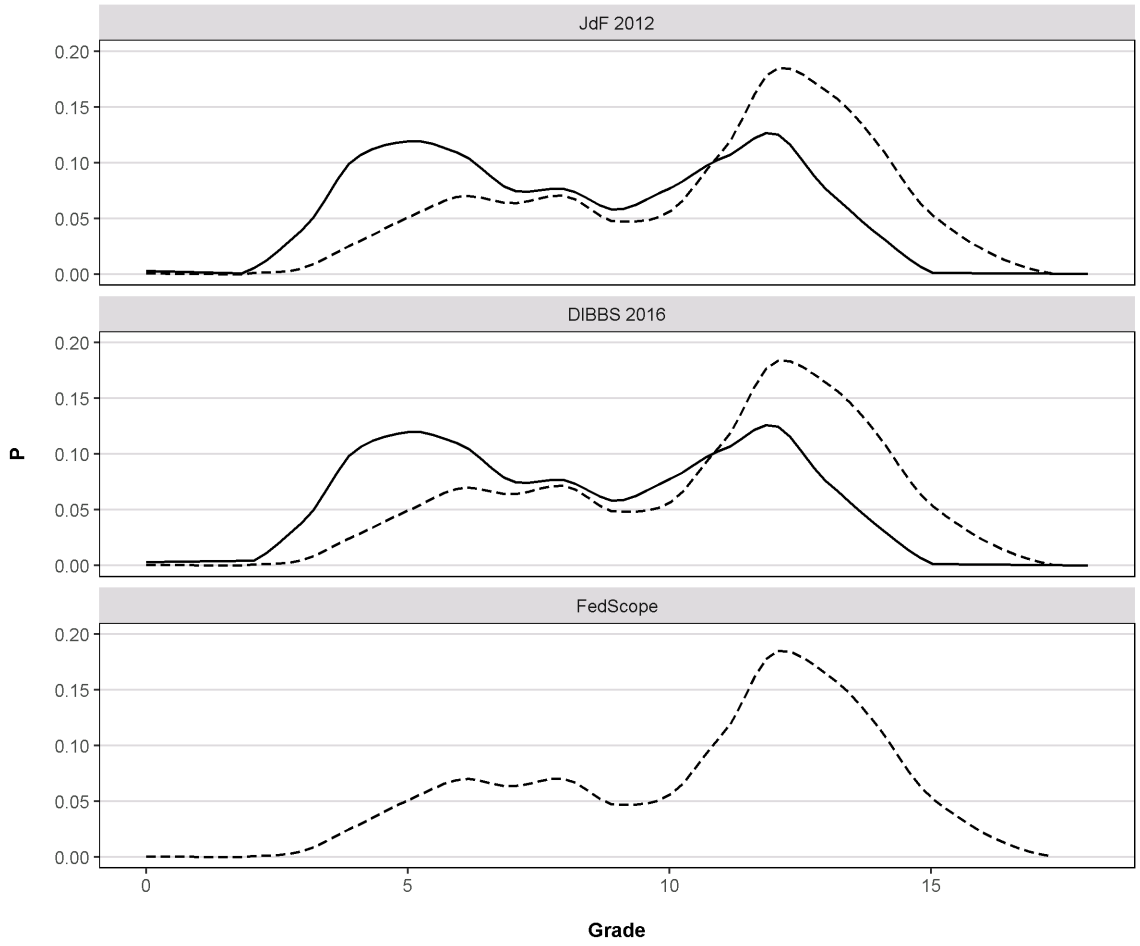


Figure 38: Change in GS grade distribution. Fiscal years 1988 (solid line) and 2011 (dashed line). Near identical distribution in synthetic and authentic data.

6.3 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 39 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.

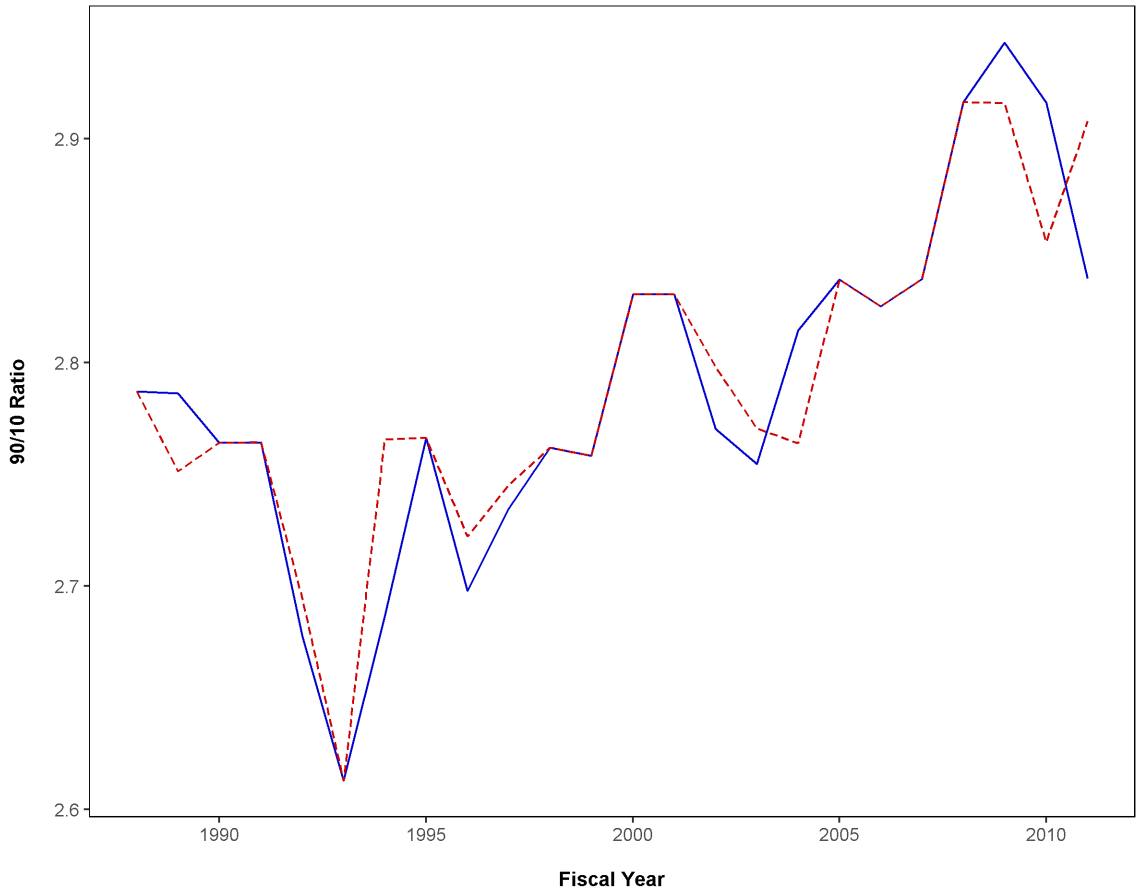


Figure 39: Ratio of 90th and 10th basic pay percentiles by year. GS pay plan, grade *leq* 15. Synthetic data dashed line, authentic data solid.

6.4 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 40 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.

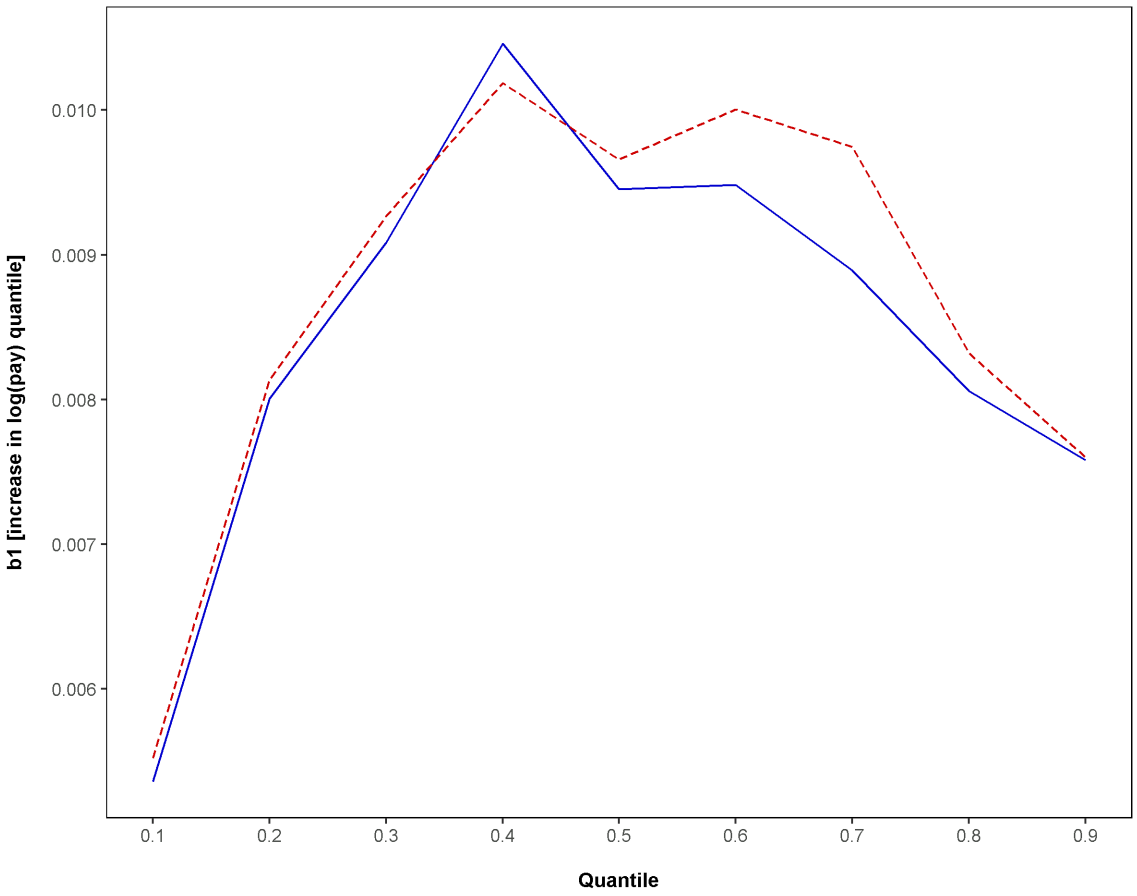


Figure 40: Coefficients (change per year) from quantile regression of log(pay) on year. GS pay plan, grade $\leq$ 15. 1988-2011. Synthetic data dashed line, authentic data solid.

6.5 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 41 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.

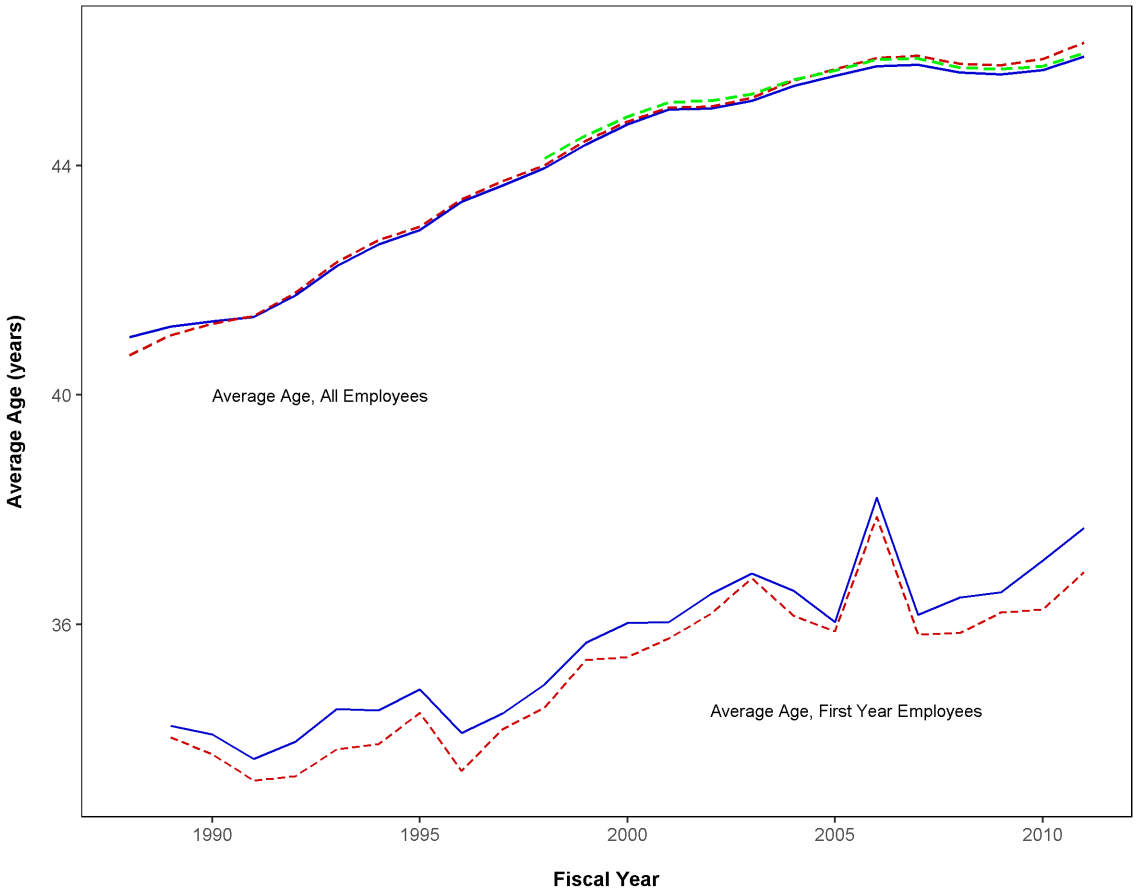


Figure 41: Change in mean age of all and first year federal employees. 1988-2011. Synthetic data dashed line, authentic data solid. Authentic trends accurately reflected in synthetic data.

6.6 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 42 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.

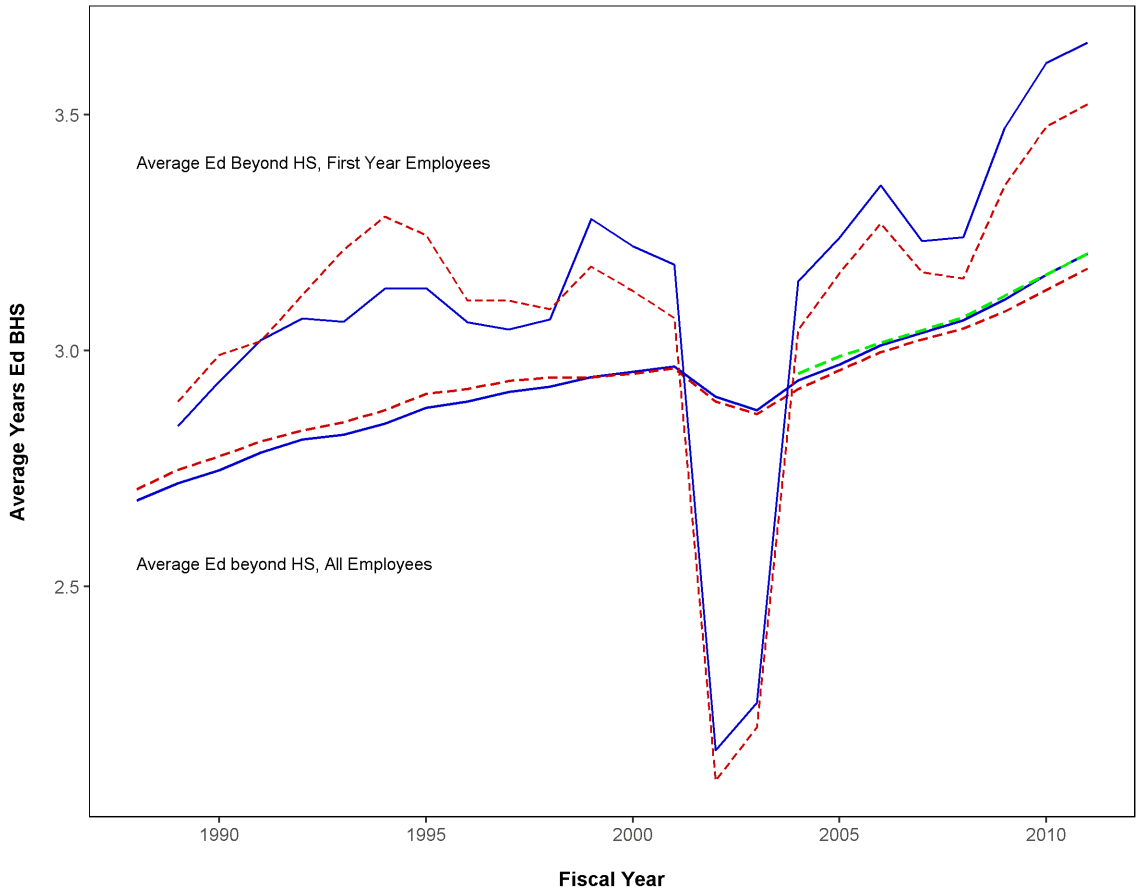


Figure 42: Change in mean years of education for all and first year federal employees. 1988-2011. Synthetic data dashed line, authentic data solid. Authentic trends accurately reflected in synthetic data.

6.7 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 43 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.

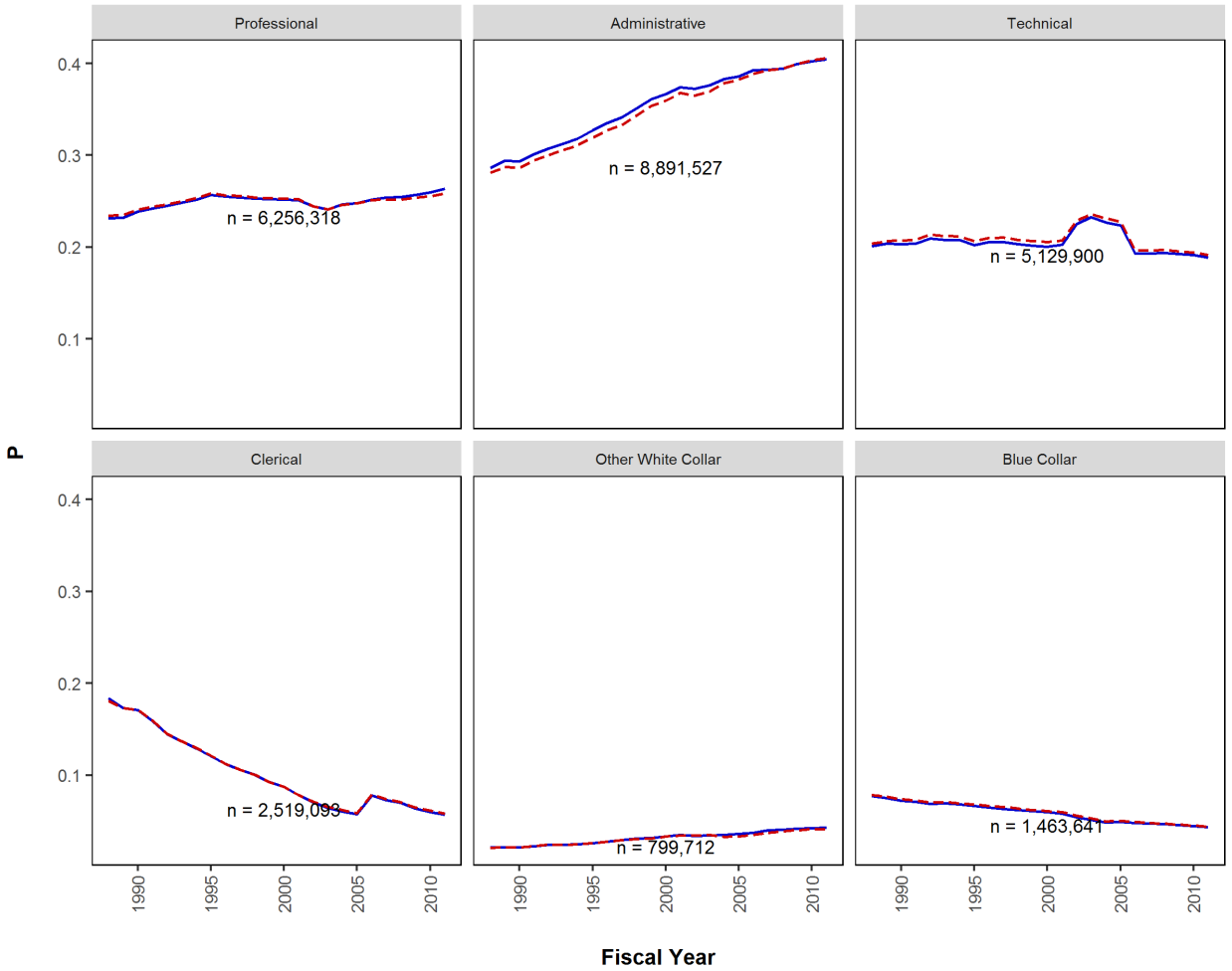


Figure 43: Change in structure of occupational categories in the U.S. federal government. 1988-2011. Synthetic data dashed line, authentic data solid. Authentic trends accurately reflected in synthetic data.

6.8 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 44 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.

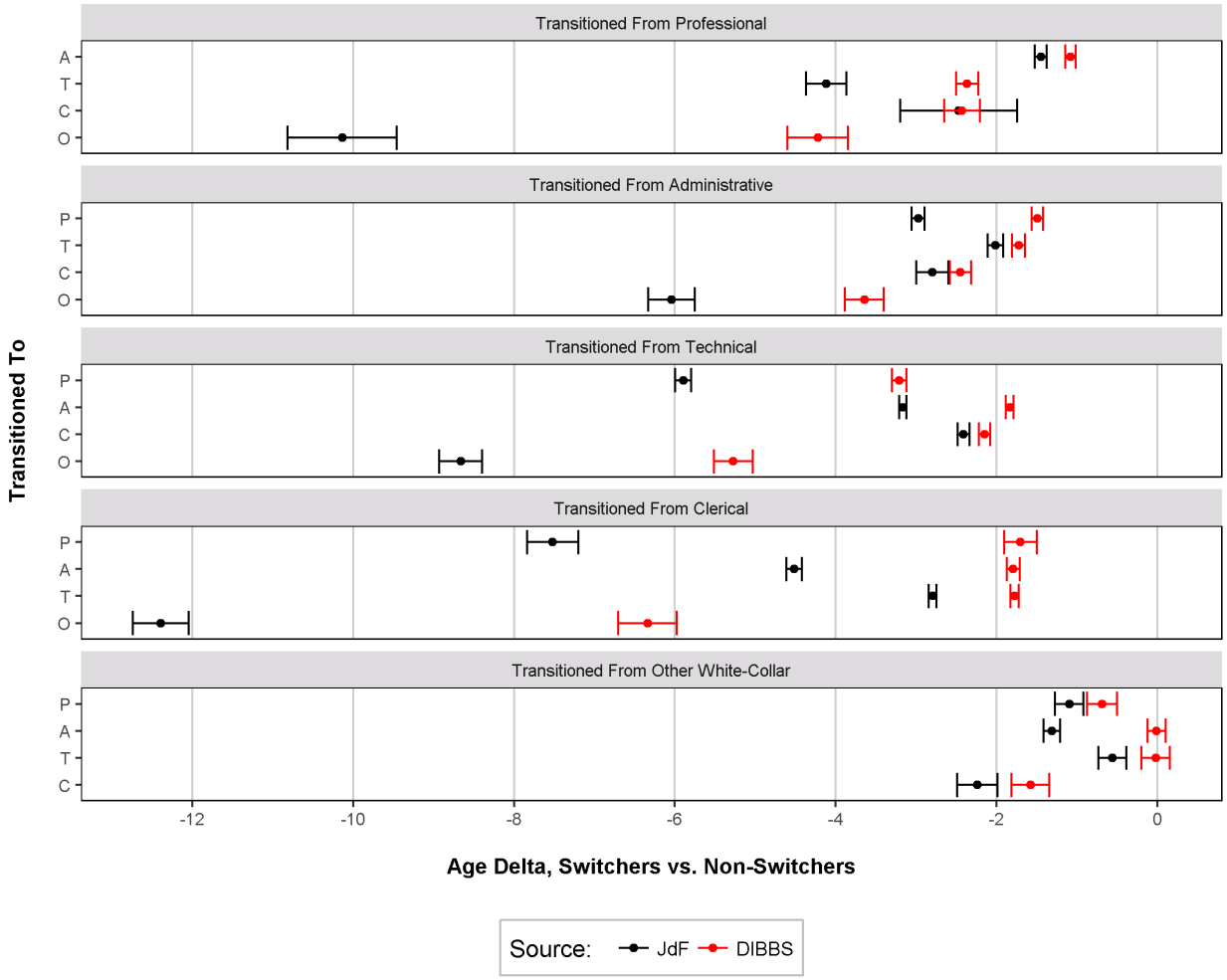


Figure 44: Mean difference in age between employees with change in occupational category and those who they join and who remained in their occupation. Non-positive means observed in both data sets.

6.9 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 45 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.

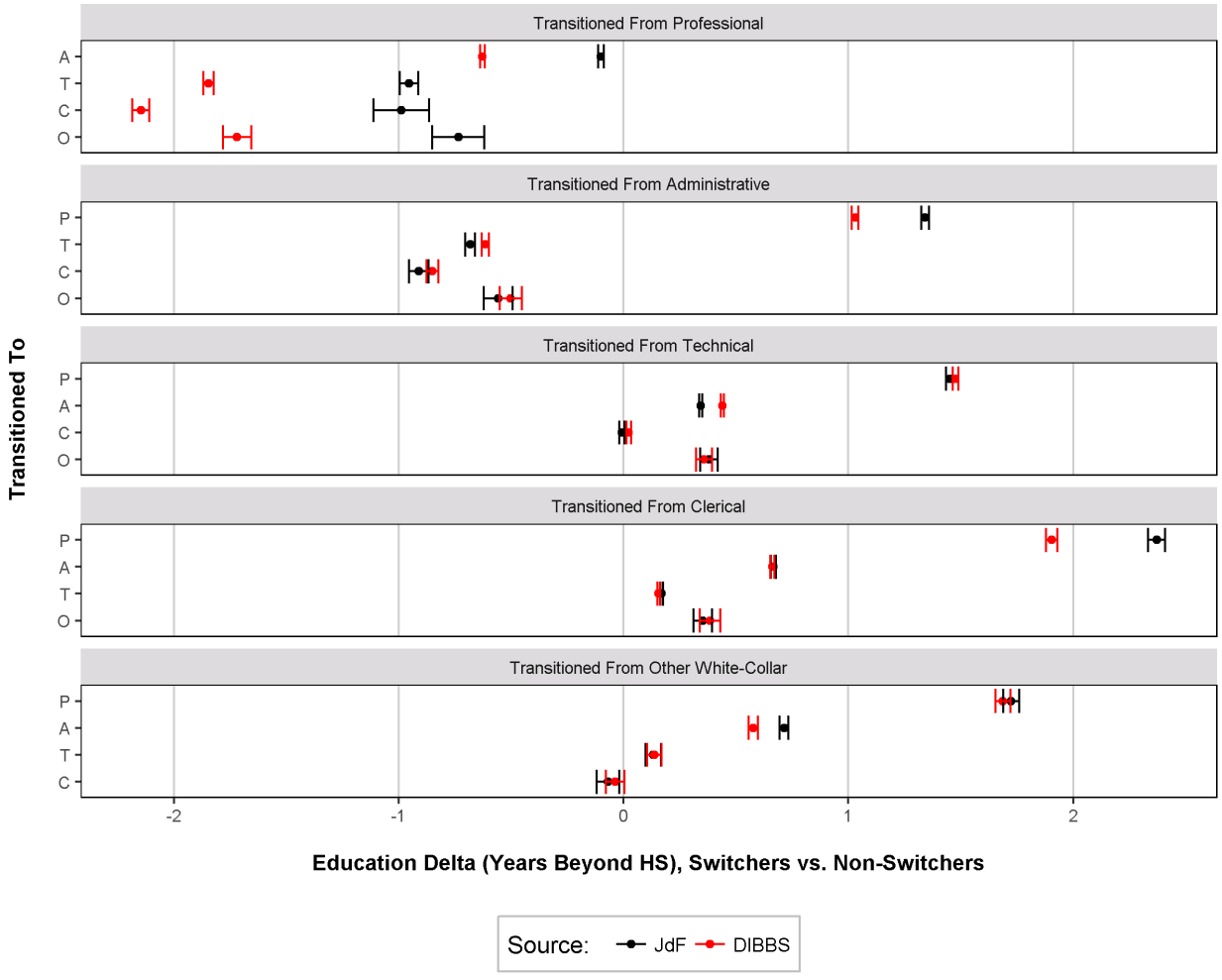


Figure 45: Mean difference in years of education between employees with change in occupational category and those who they join and who remained in their occupation. Positive means associated with categories T, C, and O, negative means associated with categories P and A.

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`.