Total Variability Measures for the LEHD Quarterly Workforce Indicators

June 10, 2019

Kevin L. McKinney, Andrew S. Green, Lars Vilhuber, and John M. Abowd

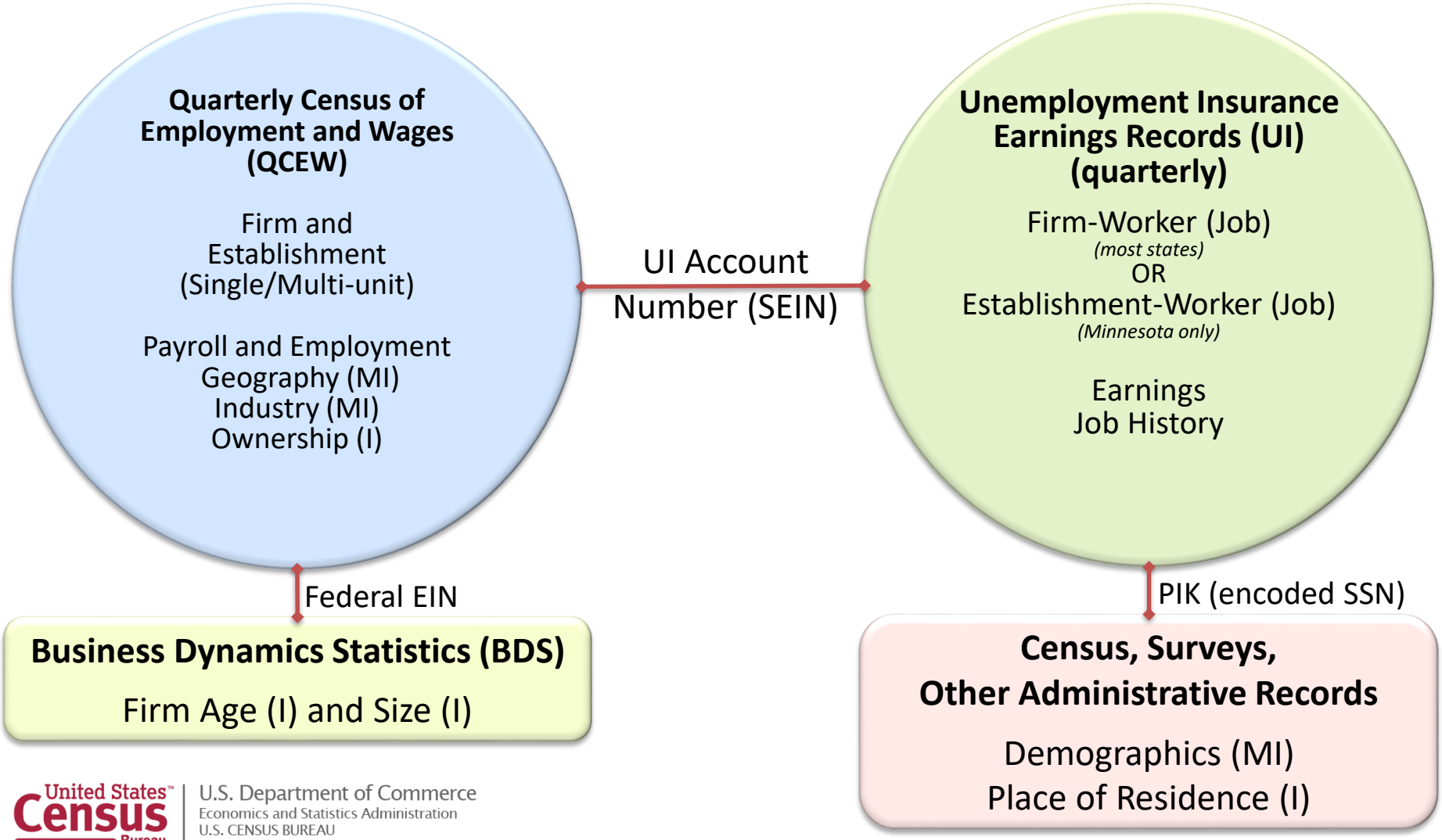
Acknowledgements and Disclaimer

- This research was conducted while all four authors were supported by the U.S. Census Bureau. Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the Census Bureau or any other research sponsor. All results have been reviewed to ensure that no confidential information is disclosed.

Background

- The Longitudinal Employer Household Dynamics Program (LEHD) at the U.S. Census Bureau uses job-level administrative data to produce detailed quarterly statistics on employment and earnings
- Data is missing for many tabulation characteristics. We complete the data using multiple imputation.
- To limit disclosure, we use multiplicative input data noise infusion
- We estimate the additional variability due to both imputation and noise infusion.

LEHD Data Sources



QWI Measures

- We evaluate five major QWI indicators.
  - Emp (M) – Number of jobs with positive earnings in the current quarter
  - Beginning of Quarter Emp (B) – Jobs with positive earnings at the same establishment in the previous and current quarter
  - Full Quarter Emp (F) – Jobs with positive earnings at the same establishment in the previous, current, and subsequent quarter
  - Average Earnings (ZW_3) – Average Earnings of F jobs
  - Payroll (W1) – Total earnings at all M jobs

QWI Tabulations

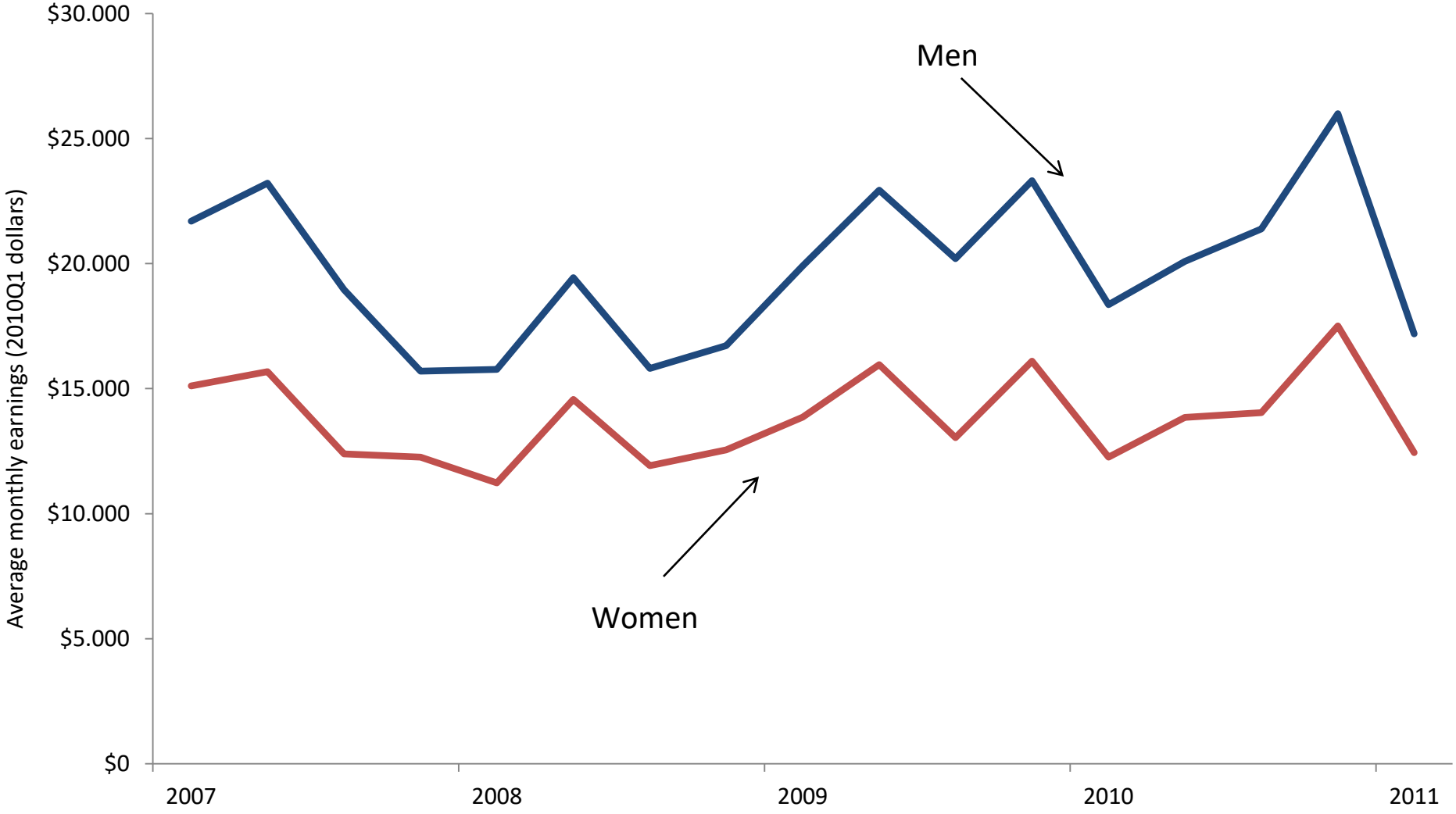
- We evaluate total variability for the following tabulations (state*year*quarter is implied)
  - Worker characteristics:
    - Age group by gender (~5% missing)
    - Race by ethnicity (~20% missing)
    - Gender by education (~80% missing)
  - We also interact each of the above tabulations with county by industry (NAICS sector) (<1% missing)

QWI Tabulations (cont.)

- Although the LEHD program is national, to reduce the computational burden we produce total variability estimates for the following 12 states: AK, DC, DE, HI, KY, ND, NH, RI, SD, VT, WV, and WY
- The tabulations are very detailed with a large number of small cells. For example, below is the cell size distribution for the measure F tabulated by state, year, quarter, industry, county, gender, and education

Cell Size	Number of Cells	Distribution
1-2	627,027	19%
3-9	844,533	25%
10-99	1,334,266	40%
100-999	468,974	14%
1000+	61,744	2%

Average monthly earnings for workers in Santa Clara County, CA in the Information Sector with a college degree 2007-2011, by gender



Note: Information is NAICS sector 51. Earnings adjusted using the CPI-U.

Major Sources of Variability

- **Unit Non-Response/Coverage Differences**
  - UI data is received and matched with the BLS' QCEW data to create a unified list frame
  - We approach the UI data as a large random sample from the integrated UI/QCEW list frame. The frame represents almost the entire population of jobs.
  - Each quarter, UI employment is at least 90% (many states are much higher, 98% or more) of the total UI/QCEW employment.
  - Weights are created so that the UI totals match the sum of state year quarter sector(private/not private) UI/QCEW employment.
- **Item Non-Response**
  - Missing tabulation characteristics (firm and/or worker) are completed using multiple-imputation
- **Disclosure Avoidance (Multiplicative Noise Infusion)**
  - Noise infusion factors are created for each establishment. We never tabulate the actual reported value for release.

Total Variability Analysis

- We use the Rubin (1987) multiple imputation approach to estimate total variability
- Within Variance
  - No sampling error, but we do have undercoverage
  - Due to the relatively large ”sample”, the median FPC over all table cells is about 0.022.
- Between Variance
  - Imputation of missing tabulation characteristics
  - Noise infusion of input data

Estimating the Variability due to Characteristic Imputation

- We create $l = 1, \ldots, L = 10$ input datasets (implicates). For each implicate l we take new draws from the posterior predictive distribution for all of the imputation models (age, gender, race, ethnicity, education, industry, and county) and recalculate the tables.
  - Across implicates, every record with at least one imputed characteristic is at risk of being assigned to a different table cell
- The higher the variability in our imputation models and the higher the proportion of missing data, the more likely a given record will be allocated to more than one table cell

Table Cell Estimates and the Within Implicate Variance

Let k represent a single category of a mutually exclusive combination of stratifying characteristics (a table cell). The Rubin estimate for k is:

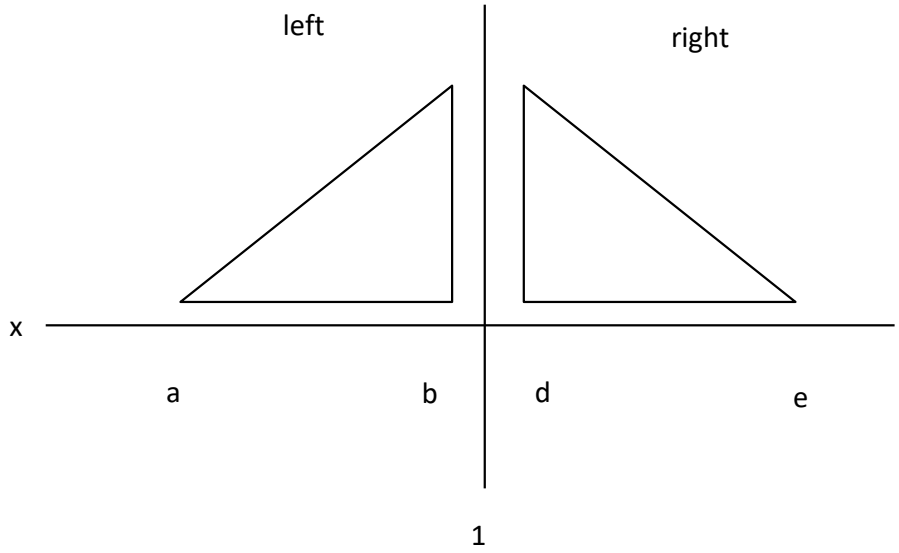
$$\bar{y}_k^* = \frac{1}{L}\sum_{\ell=1}^{L} y_k^{(\ell)*}$$

The Rubin average within-variance:

$$\overline{vy}_k^* = \frac{1}{L}\sum_{\ell=1}^{L} vy_k^{(\ell)*}$$

Noise Infusion Distribution

Double Sided Symmetric Ramp Distribution



(1-b)=(d-1)
(b-a)=(e-d)

Estimating the Variability due to Noise Infusion

- We create an additional $l = 1, \ldots, L = 10$ input datasets (implicates)
- For each implicate l and establishment j, we draw a new noise infusion factor δ_j^l, holding constant the imputed characteristics at the $l = 1$ values
- The fewer establishments and/or the more unequal the distribution of jobs across establishments in a table cell, the higher the variance

Between Contribution to Total Variability

The between implicate variance due to imputation

$$bcy_k \frac{1}{L-1}\sum_{\ell=1}^{L}\left(y_k^{(\ell)} - \bar{y}_k\right)^2$$

The between implicate variance due to noise infusion

$$bsy_k^* = \frac{1}{L-1}\sum_{\ell=1}^{L}\left(y_k^{(\ell)*} - \bar{y}_k^*\right)^2$$

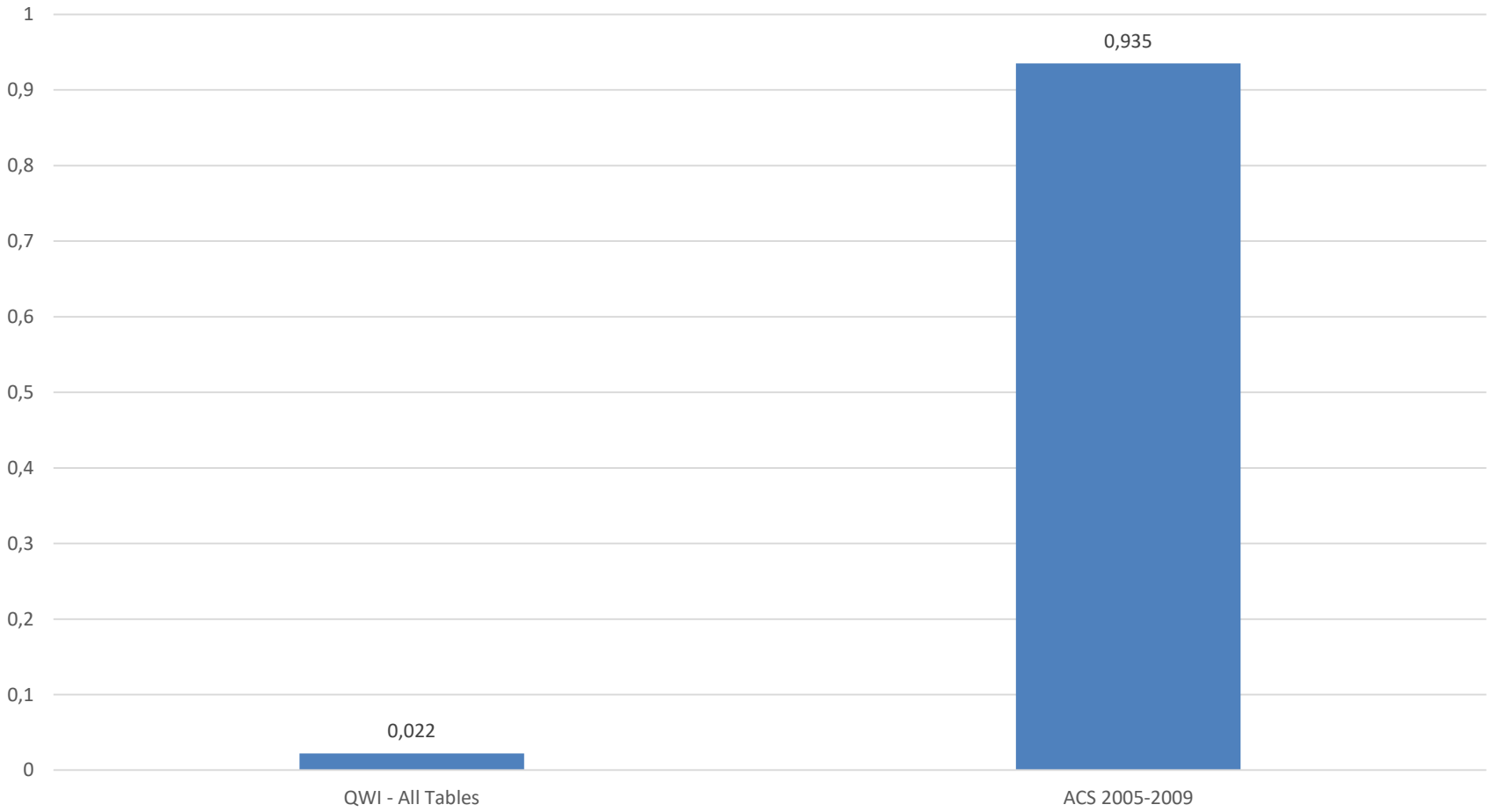
Total Variability

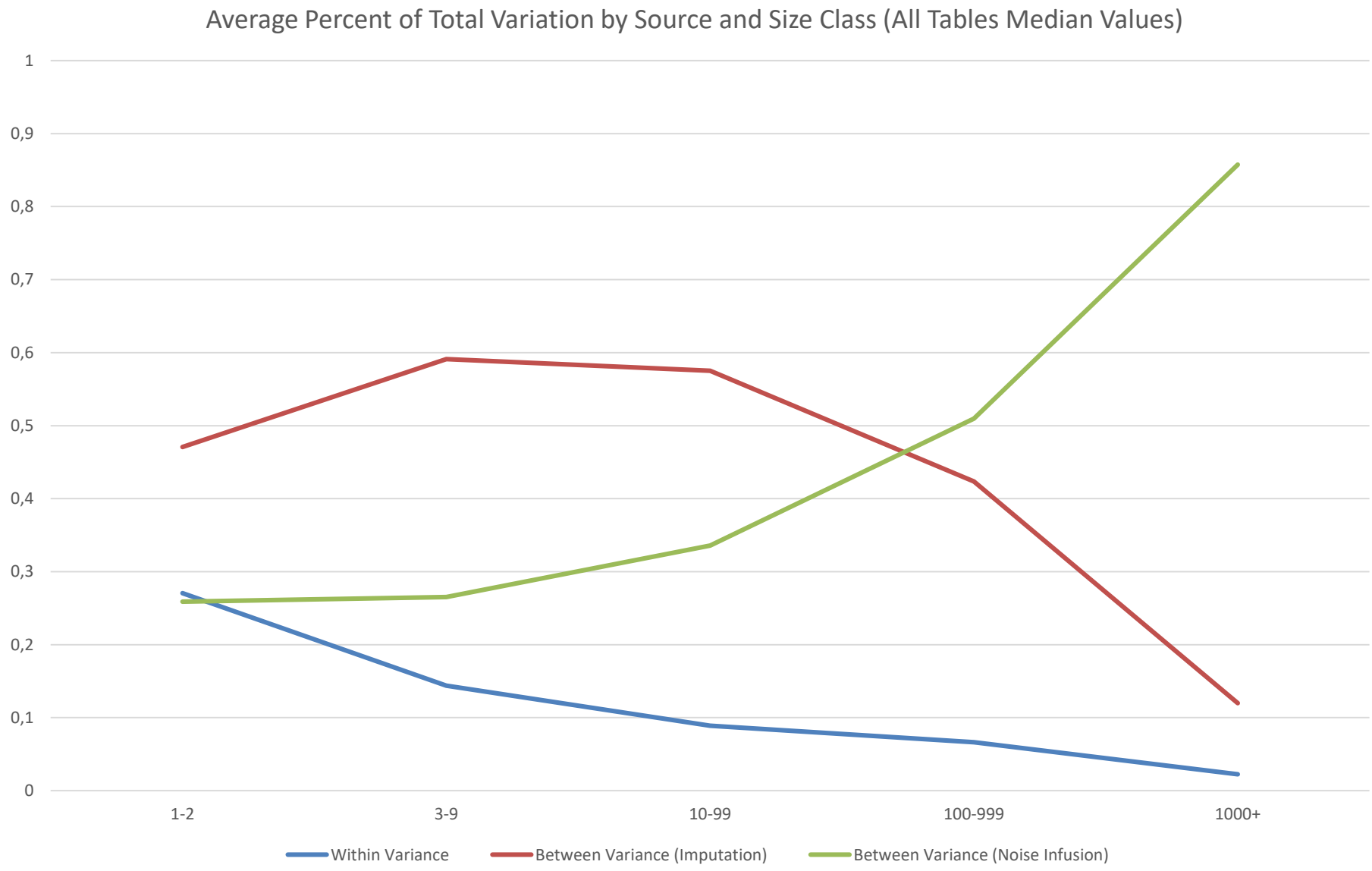
The Rubin total variance (weighted average of within and between variance components):

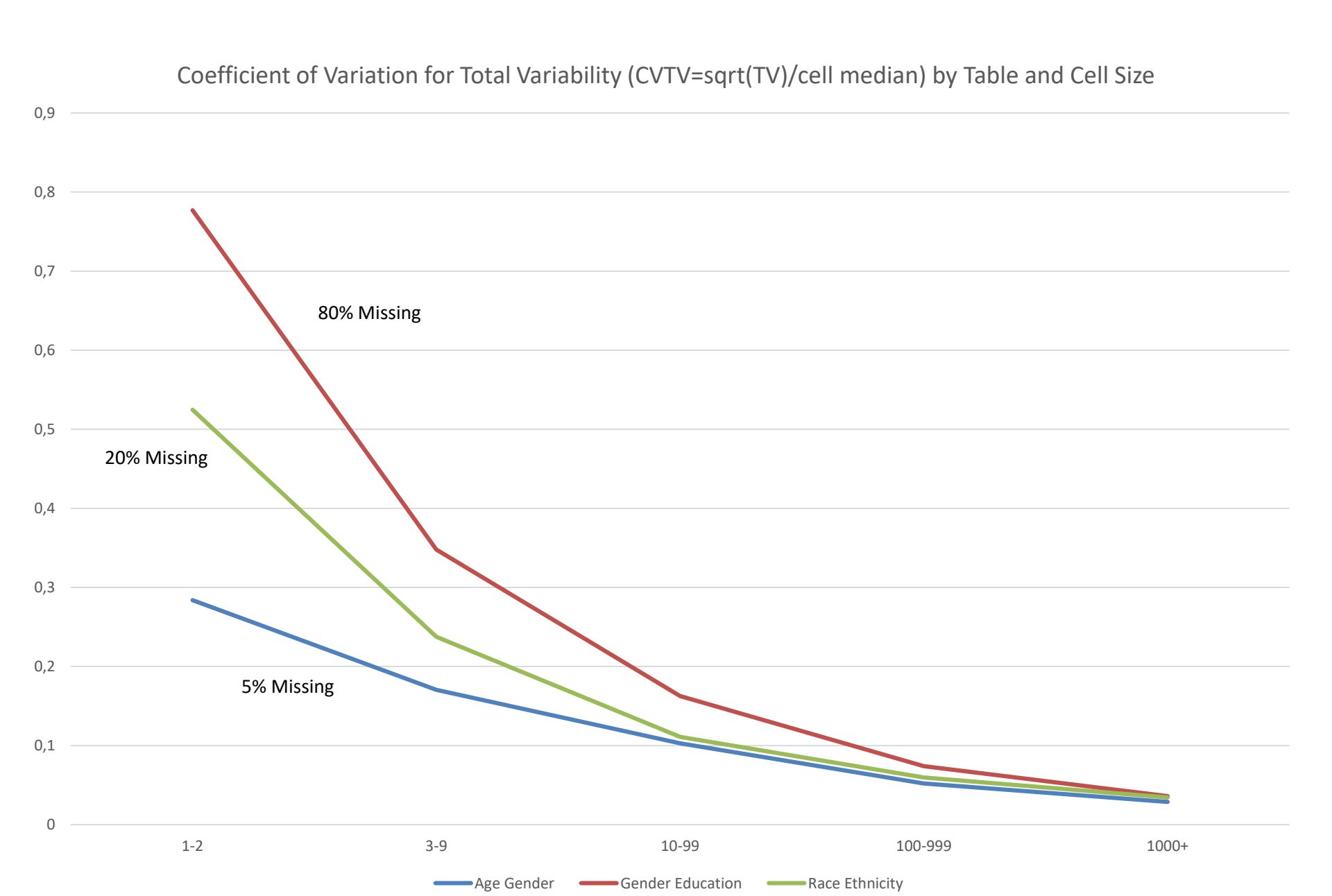
$$tvy_k^* = \overline{vy}_k^* + \frac{L+1}{L} bcy_k + \frac{L+1}{L} bsy_k^*$$

Results

Median FPC Correction




Average Percent of Total Variation by Source and Size Class (All Tables Median Values)
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
1-2
3-9
10-99
100-999
1000+
Within Variance
Between Variance (Imputation)
Between Variance (Noise Infusion)



United States Census Bureau
U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Confidence Interval Comparison

Total Variability (Median) B Gender Education				
			Confidence Interval	
Cell Size	Median Size	sqrt(TV)	Bottom	Top
1-2	1	0.7	-3.7	5.7
3-9	5	1.8	1.3	8.7
10-99	26	4.2	18.5	33.5
100-999	212	15.3	184.3	239.7
+1000	1,739	59.4	1,630.2	1,847.8

Within Variability Only (Median) B Gender Education				
			Confidence Interval	
Cell Size	Median Size	sqrt(WV)	Bottom	Top
1-2	1	0.2	-0.4	2.4
3-9	5	0.5	4.0	6.0
10-99	26	1.1	24.2	27.8
100-999	212	3.3	206.6	217.4
+1000	1,739	9.3	1,723.7	1,754.3

Conclusion

- Due to the large “sample” size the FPC is small, greatly reducing within implicate variability
- Without taking account of imputation and noise infusion the standard variability measure overstates the reliability of the data (notable for smaller cells)
- The total variability approach gives the user a more complete picture of the sources of error
- Especially important for large sample administrative data where the traditional sources of error are small
- Future – Missing reports and processing errors