Imputation in the Census of Manufactures

T. Kirk White[†]
Jerome P. Reiter**
Amil Petrin*

†U.S. Census Bureau
**Duke University
*University of Minnesota and NBER

The research in this presentation was conducted while the authors were, respectively a employee of the Census Bureau, and Special Sworn Status researchers of the U.S. Census Bureau at the Triangle Census Research Data Center and the Minnesota Census Research Data Center. Research results and conclusions expressed are those of the authors and do not necessarily reflect the views of the Census. This presentation has been screened to ensure that no confidential data are revealed.

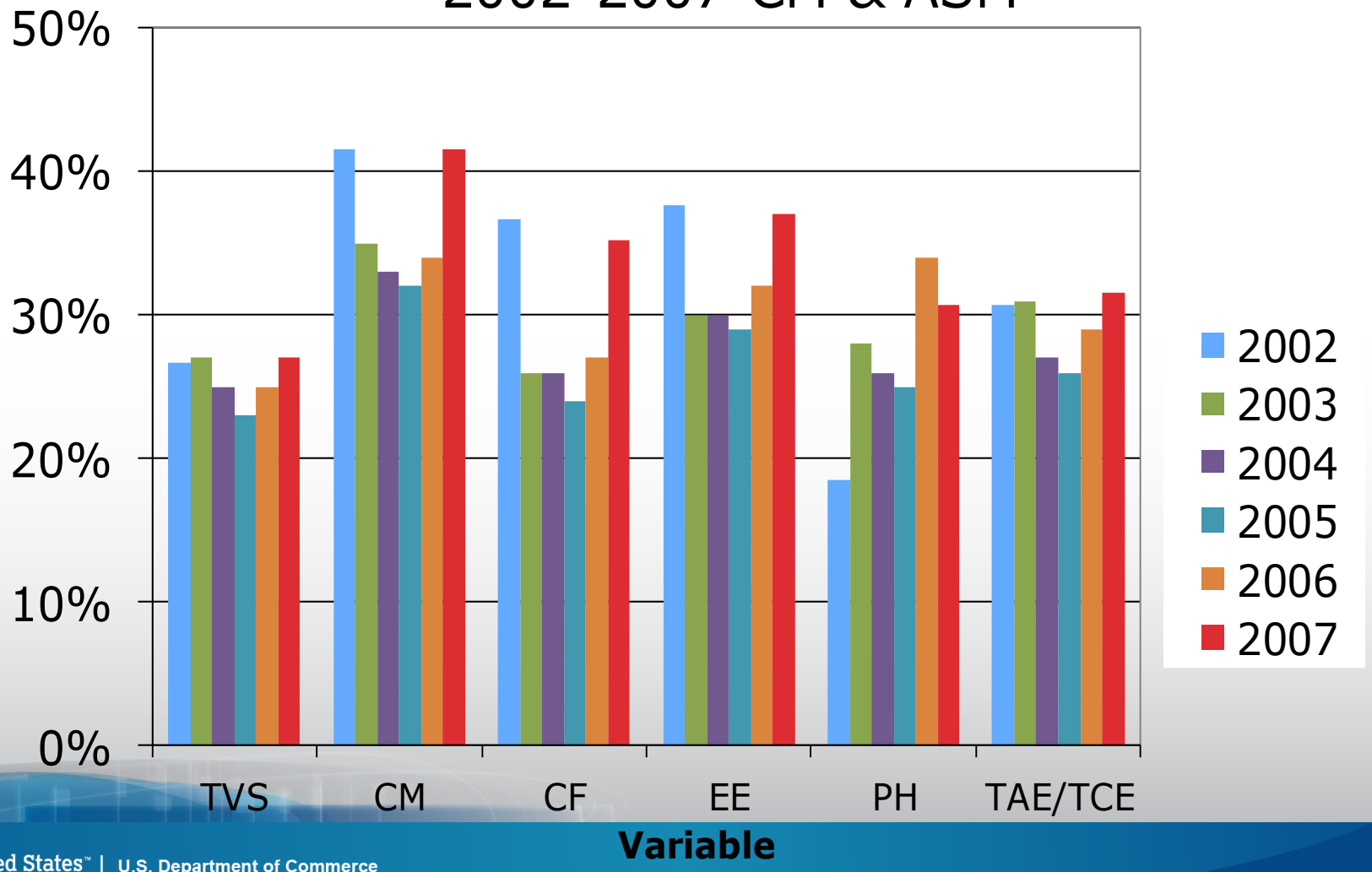
The U.S. Census of Manufactures

- Conducted every 5 years (part of Economic Census)
- Includes data for about 300,000-400,000 plants
- Includes a sample of 50,000-70,000 plants as part of the Annual Survey of Manufactures (annual panel)
- Data on revenue, employment, payroll, inventories, and production expenses are collected in all manufacturing industries (same questionnaire)
- Industry-specific data on materials used in production and specific products (about 300 different questionnaires)

Imputation in the Census of Manufactures

- Most very small plants (< 5 employees) are not sent a questionnaire – roughly one third of all plants
  - Data for these plants is imputed using administrative records data on payroll, employment, and sales
  - Most researchers (including us) ignore these plants
- For other plants, the Census Bureau (singly) imputes data for three reasons:
  - Unit non-response
  - Item non-response
  - Response data fails edit checks

Mean NAICS6 Industry Imputation Rates, Excluding "Non-mail" Sample, 2002-2007 CM & ASM

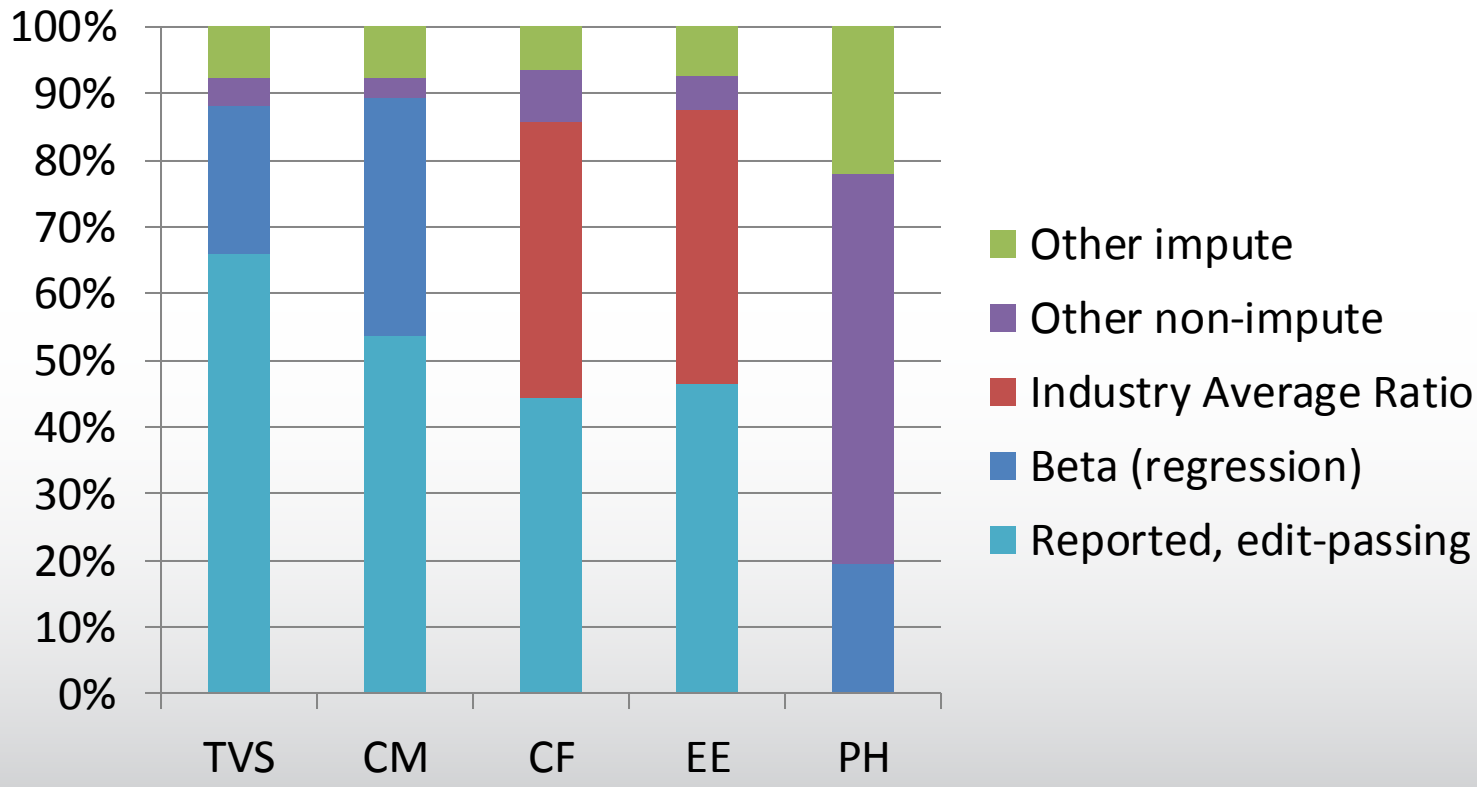


Imputation Methods Used in the Census of Manufactures

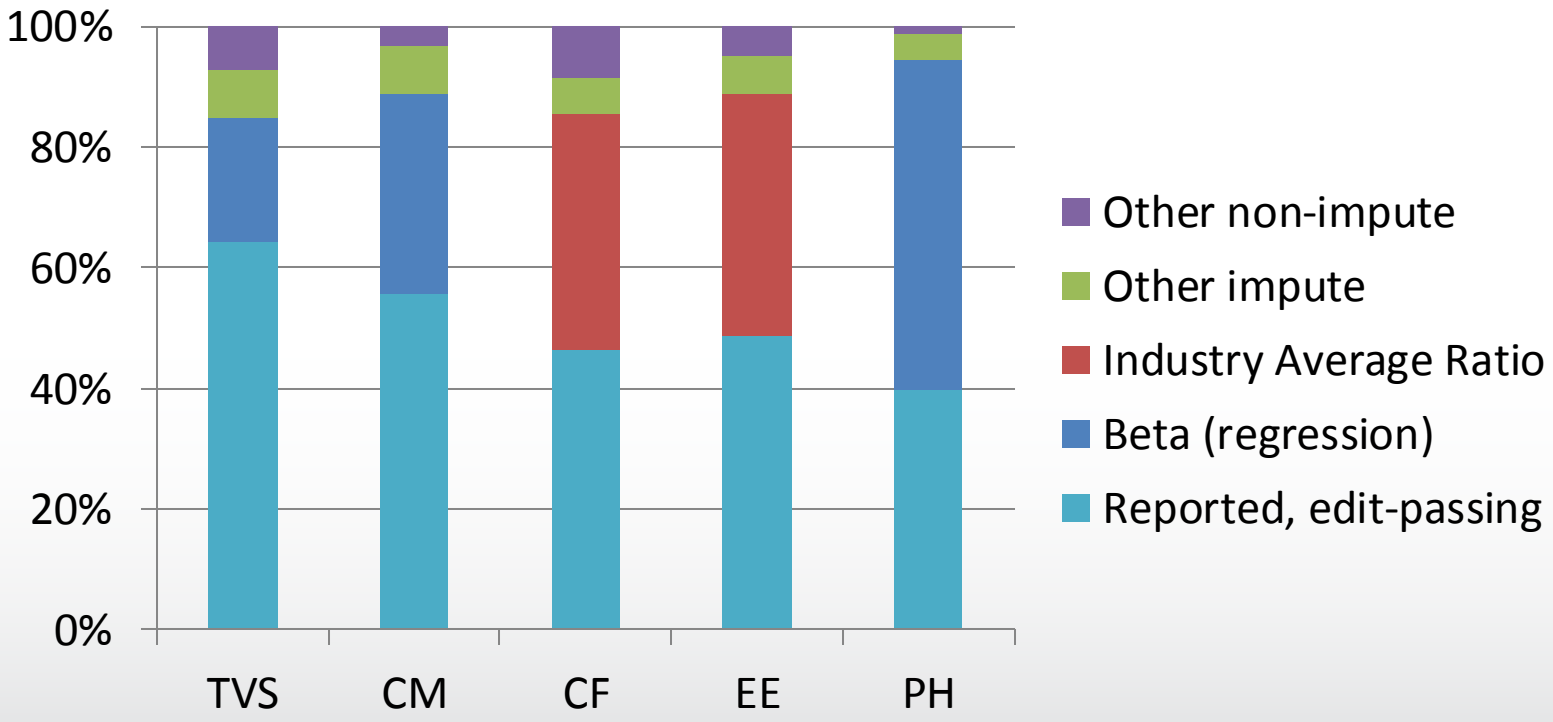
- Industry-specific regression models:
  - $Y_{ijt}^{impute} = B_j X_{ijt}$

    or
  - $Y_{ijt}^{impute} = B_{1j} X_{ijt} + B_{2j} Y_{ij,t-1} + B_{3j} X_{ijt-1}$
- Industry Average Ratio: $Y_{ij}^{impute} = X_{ij} \underline{(Y/X)}_j$
- Substitution from current-year or prior-year administrative records for same plant (payroll, employment, sales)
- Prior year reported data for same plant (ASM sample only)
- Logical relationships for same plant (e.g., sum of details)
- Other methods (used infrequently)

Reported vs. Imputation Methods, 2002 Census of Manufactures



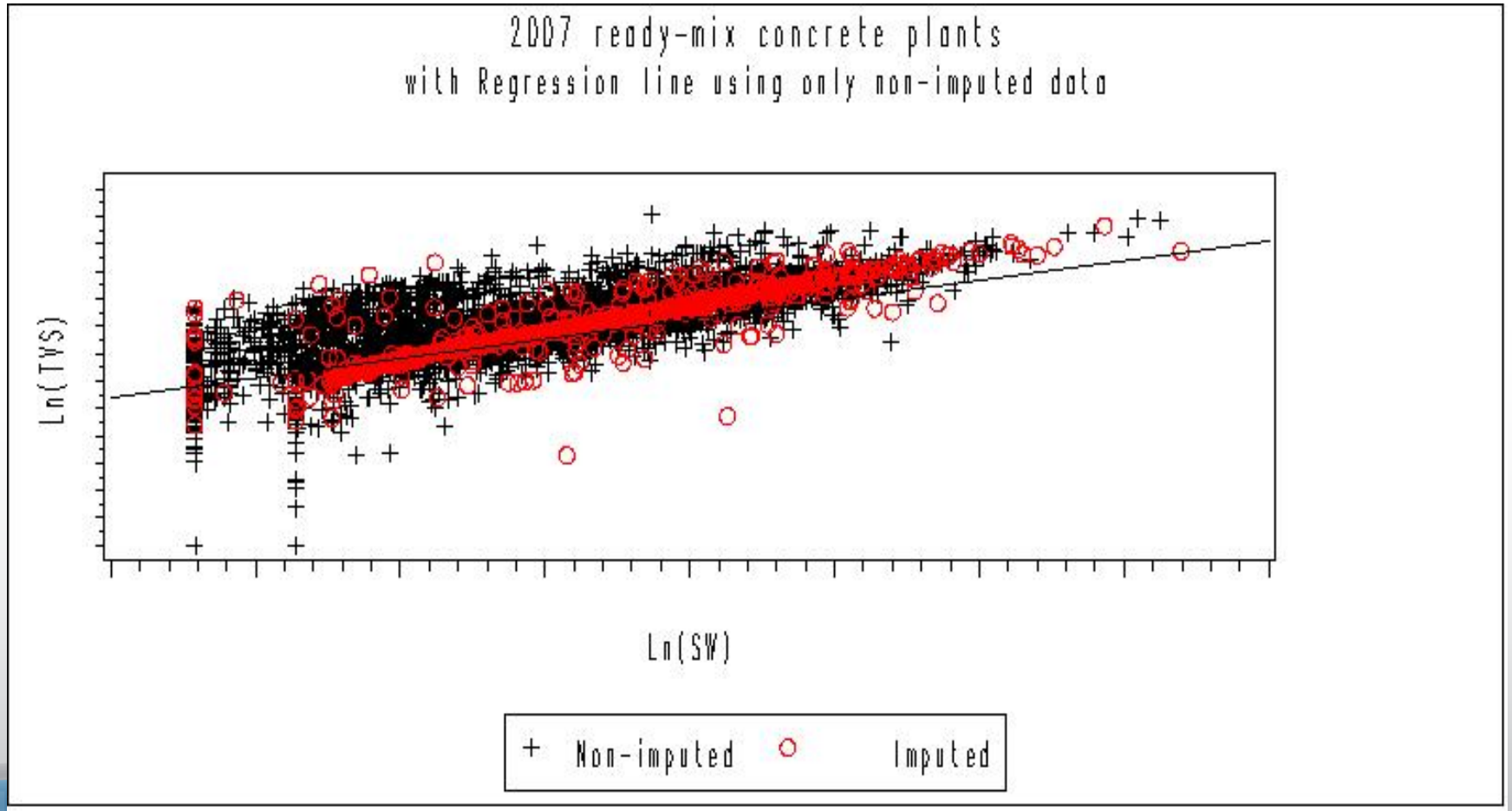
Reported vs. Imputation Methods, 2007 Census of Manufactures



Measuring Effects of Imputations on Within-Industry Dispersion in the CM

- For each of 458 industries, each year (2002-2007), and each key input variable X, we calculate the within-industry Interquartile Range (IQR) of X over Total Value of Shipments:
  1. When X is imputed
  2. When X is not imputed
- For most industries there is much less within-industry dispersion in the conditional distributions of the imputed data than in the non-imputed data

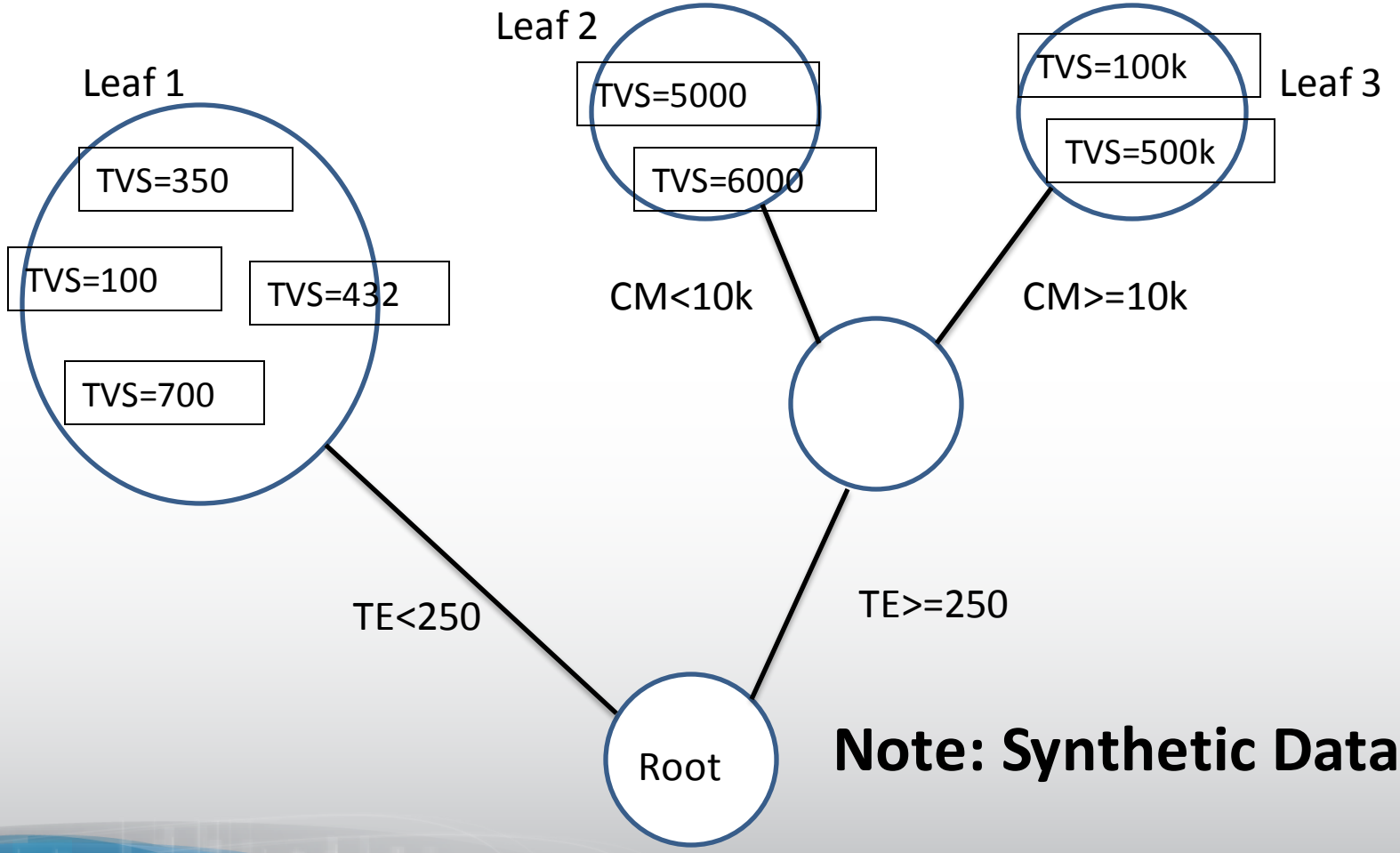
Within-Industry Dispersion in a Conditional Distribution, Imputed vs. Non-imputed Data



Imputations using Sequential Classification and Regression Trees (CART)

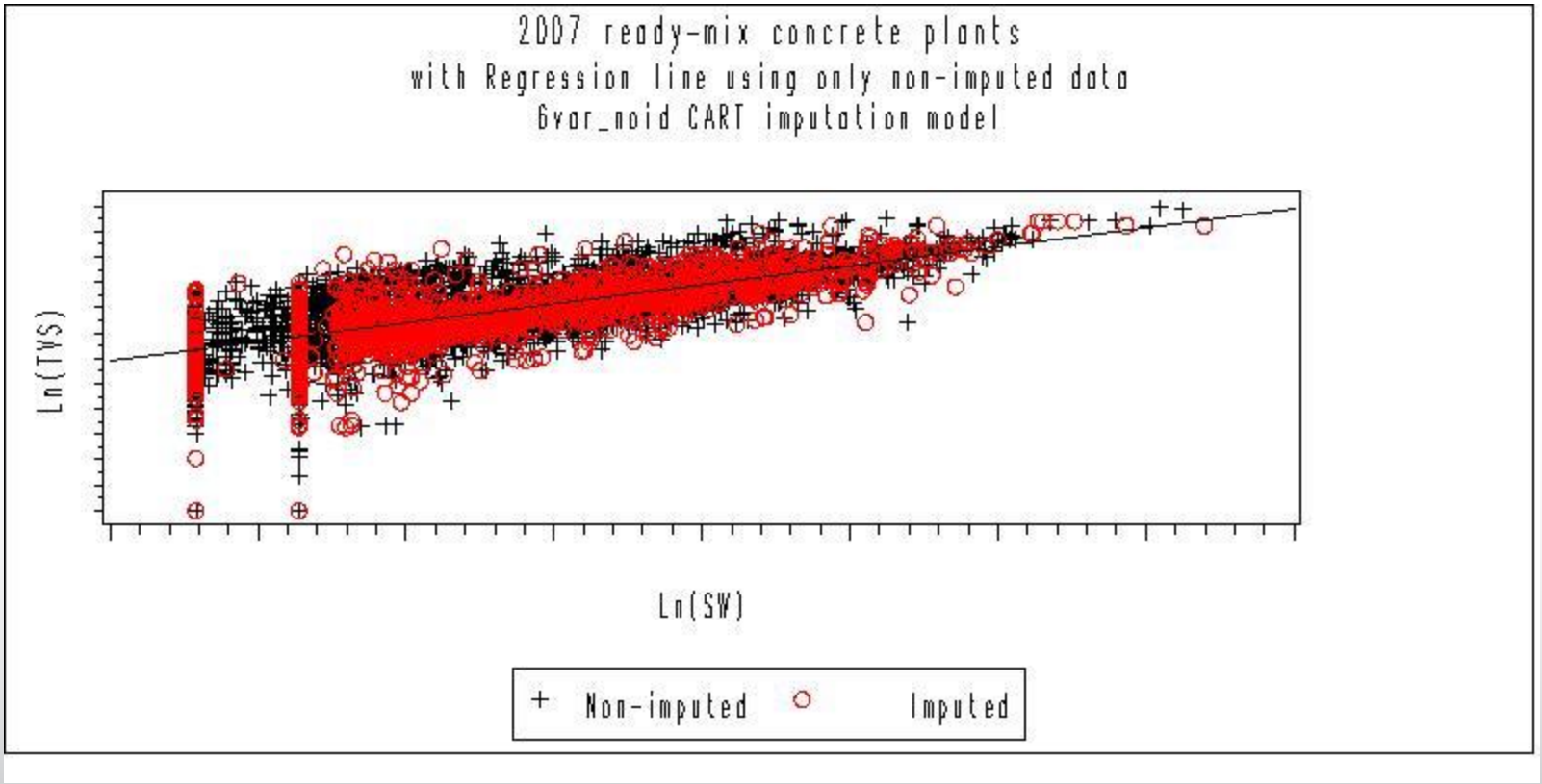
- We replace the Census Bureau's imputations in the Census of Manufactures with multiple imputations using the sequential CART method (Burgette and Reiter 2010)
- Advantages:
  - Works well with highly skewed distributions like we see in CM data
  - Fits interactions and non-linear relationships without parametric assumptions
  - Generates appropriately dispersed imputations
  - Allows for valid variance estimation (appropriate standard errors)
- Disadvantage:
  - More computationally-intensive than Census Bureau's methods

A Classification and Regression Tree (CART)

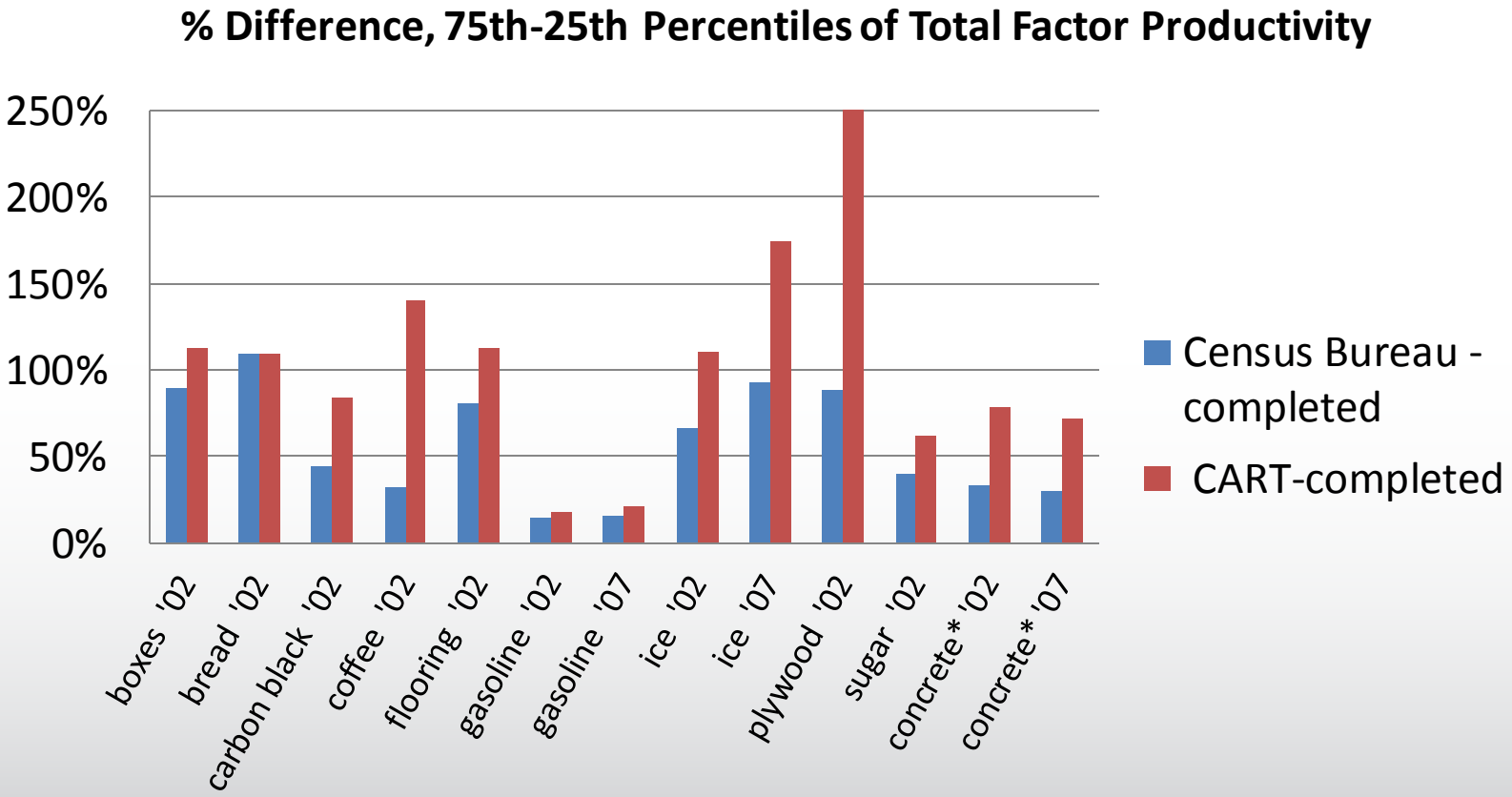


Note: Synthetic Data

Within-Industry Dispersion in a Conditional Distribution, CART-imputed vs. Non-imputed Data



Within-Industry Productivity Dispersion Bureau-completed vs. CART-completed Data



Conclusions

- Significant percentages of observations for key variables in the Census of Manufactures are imputed
- Census Bureau's current imputation methods in the CM reduce within-industry dispersion
- The sequential CART method does a better job of approximating the observed conditional distributions
- Still much research to be done
  - Use more administrative records data, lagged values
  - Computational time is an issue in a production environment