Comparison of Hot Deck and Multiple Imputation Methods Using Simulations for HCSDB Data

Donsig Jang, Amang Sukasih, Xiaojing Lin
Mathematica Policy Research, Inc.

Thomas V. Williams
TRICARE Management Activity

The 7th International Conference on Health Policy Statistics
January 18, 2008

MATHEMATICA
Policy Research, Inc.

Disclaimer

The opinions and assertions are those of the authors and do not reflect the views or policies of the DoD or the US government

MATHEMATICA
Policy Research, Inc.

Goal of the Study

- To determine an imputation method to be used for Health Care Survey of DoD Beneficiaries (HCSDB) item nonresponse

- Impute or not to impute?

  OMB Guideline (2006):

  - If IRR < 70%, conduct item nonresponse bias analysis
  - Unlike unit nonresponse, nonresponse adjustment like imputation is NOT required for item nonresponse

MATHEMATICA
Policy Research, Inc.

Why Imputation for Item Nonresponse?

- No imputation
  - Cumulative item nonresponse rates can be prohibitively large
  - Substantial reduction of valid data for multivariate analysis
- Imputation
  - Provide a complete dataset for analysis
  - Fill in troublesome missing data for data users
  - Prevent multivariate software packages from ignoring observations with missing data
  - Hopefully reduce nonresponse bias

HCSDB

- Quarterly survey on the military health system's beneficiaries' opinions about their DoD health care benefits
- Administered by TRICARE Management Activity (TMA)
- A stratified random sample: partitions beneficiaries based on *enrollment type*, *beneficiary group*, and *geographic areas*
- Item nonresponse rates of HCSDB key items:
  - Between 2% – 13%: Moderate
  - Similar nonresponse rates across quarters

MATHEMATICA
Policy Research, Inc.

HCSDB

- MPR's research on the need of imputation:
  - Ready to implement imputation to all survey items?
  - Which one—hot-deck or multiple imputation?
  - Compare through empirical investigation, including simulation works
- Imputation methods considered:
  - For single imputation, within-class unweighted sequential hot deck method
  - For multiple imputation, sequential regression multivariate imputation technique (SRMI)

Within-class Unweighted Sequential Hot-deck

- In-house SAS macros developed based on Carlson, Cox, and Bandeh (1995)

- HCSDB hot-deck:
  - 53 variable candidates for covariates, including survey items, demographic and sampling variables
  - modeling to choose control variables (α = 10%)

Sequential Regression Multivariate Imputation (SRMI)

- Raghunathan, Lepkowski, Hoewyk, and Solenberger (2001)
- $Y_1, Y_2, \ldots, Y_k$ denote *k* survey items with missing values
- *X* denotes a set of fully observed covariates

$$f(Y_1, Y_2, \cdots, Y_k \mid X, \theta_1, \theta_2, \cdots, \theta_k) =$$
$$f(Y_1 \mid X, \theta_1) f(Y_2 \mid X, Y_1, \theta_2) \cdots f(Y_k \mid X, Y_1, Y_2, \cdots, Y_{k-1}, \theta_k)$$

SRMI

- Each $f(Y_i|\bullet)$ is modeled through a regression model appropriate for the variable type of Y_i
- Imputation for missing Y_i is then drawn from a posterior predictive distribution of Y_i through the regression model
- Each imputation is done iteratively through *c* rounds

Multiple Imputation

Results in *M* datasets

$$\hat{\theta}_M = \frac{1}{M}\sum_{i=1}^{M}\hat{\theta}_i$$

$$\mathrm{var}(\hat{\theta}_M) = W_M + \left(1 + \frac{1}{M}\right)B_M$$

where

$\hat{\theta}_M$ = estimate based on multiply imputed data

$\hat{\theta}_i$ = estimate based on the i-th imputed value

$\mathrm{var}(\cdot)$ = variance of the esimate

Multiple Imputation

Variance component

$$W_M = \frac{1}{M}\sum_{i=1}^{M} \mathrm{var}(\hat{\theta}_i) = \text{ within-imputation variance component,}$$

$$B_M = \frac{1}{M-1}\sum_{i=1}^{M}\left(\hat{\theta}_i - \hat{\theta}_M\right)^2 = \text{ between-imputation variance component.}$$

MI Implementation: IVEware

- Raghunathan, Solenberger, and Van Hoewyk (2002)
- SAS-callable software application
- Proper regression model for Y_j
- Model-building algorithm, variable interactions, stepwise method, specified minimum marginal R^2
- Accounts for complex data structures from a complex questionnaire, e.g. skip pattern/restriction and bounds

MI Implementation using IVEware (Continued)

- Among predictors were key sampling and demographic variables
- Used survey weights and weighting adjustment cells as predictors
- Computation for categorical variables with many categories can take a great deal of computational time

Simulation Study

- Use of Quarter 1 of 2003 HCSDB response data
- Generate missing values with missing rates of 2 – 13% under three different missing mechanisms (MCAR, MAR, NMAR)
- Impute using hotdeck and multiple imputations
- Two variance estimators under hot-deck imputed data
  - Assume the imputed value as if it were a true value
    - TS method
    - Jackknife
- Compare estimates and MSE (bias and variance)

MATHEMATICA
Policy Research, Inc.

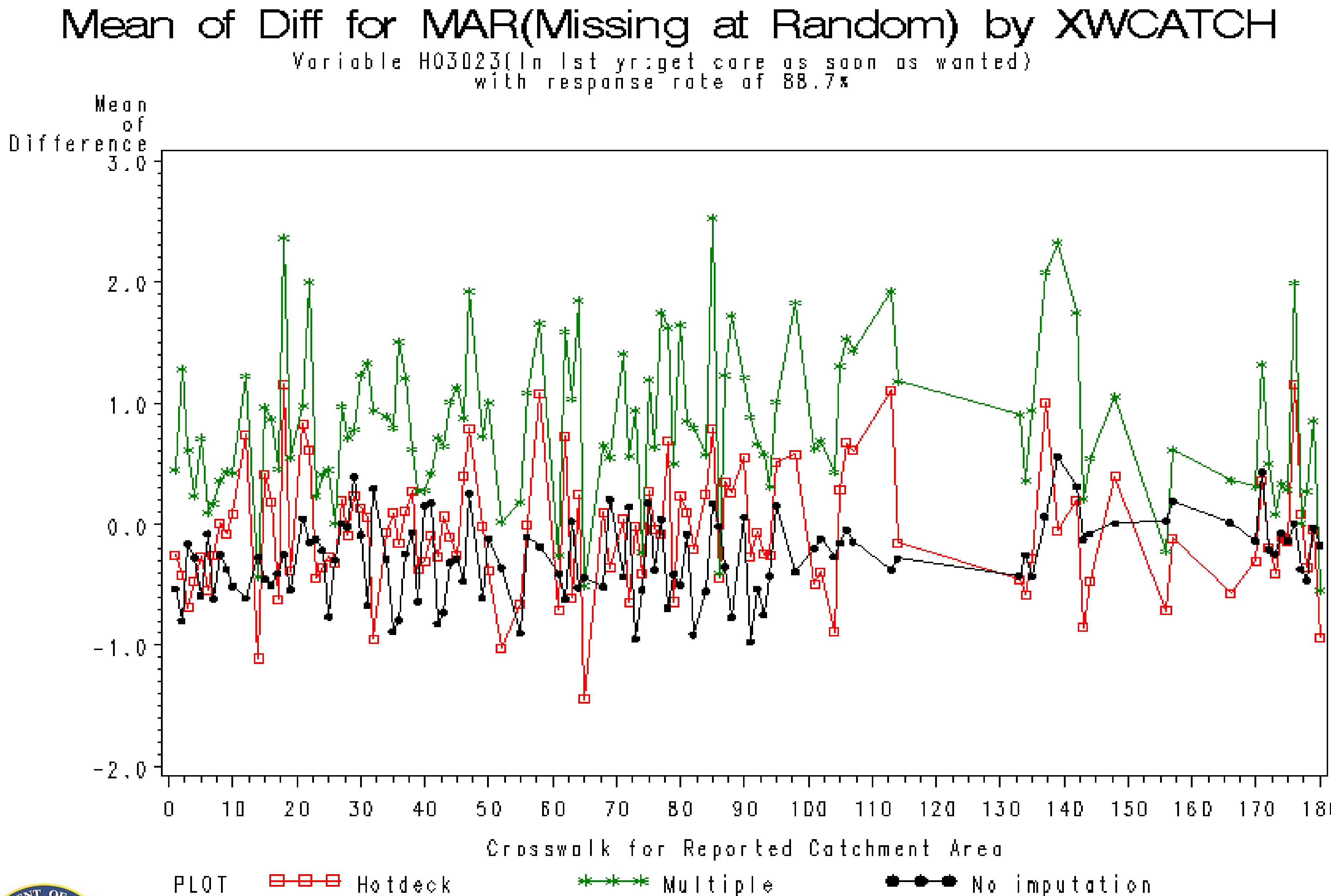
Results for Selected Variable

Variable H03023 (88.7% response rate)

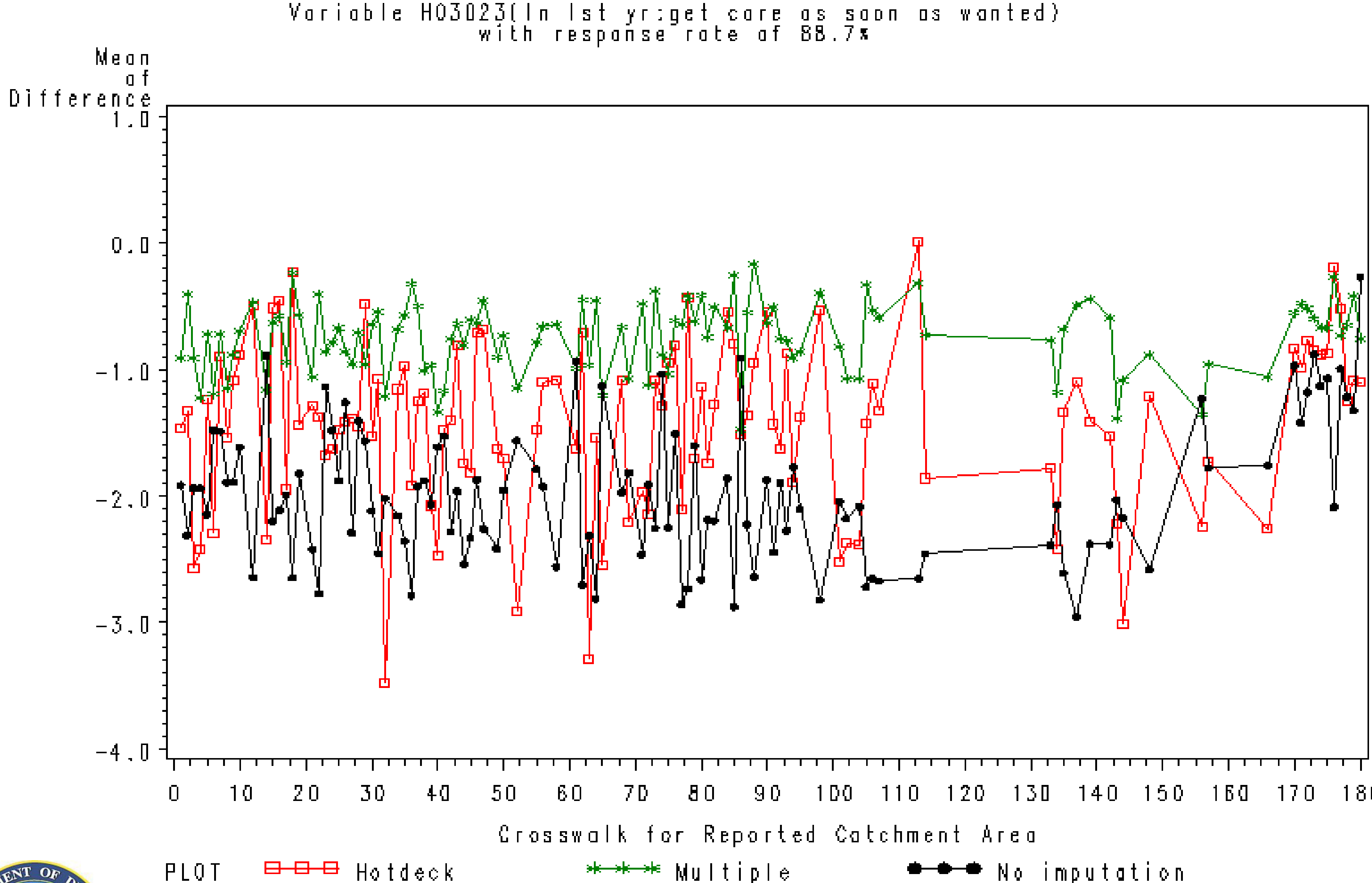
$$\textbf{Mean Diff }(p) = \frac{1}{R}\sum_{r=1}^{R}(p_r - p_0)\times 100\%$$

$$\textbf{Square-root Mean Square Diff }(p) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(p_r - p_0)^2}\times 100\%$$

$$\textbf{Rel.var} = \frac{\frac{1}{R}\sum_{r=1}^{R}V_r}{V_0}\times 100\%$$


Mean of Diff for MAR(Missing at Random) by XWCATCH
Variable H03023(In 1st yr:get care as soon as wanted)
with response rate of 88.7%
Mean of Difference
3.0
2.0
1.0
0.0
-1.0
-2.0
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180
Crosswalk for Reported Catchment Area
PLOT
Hotdeck
Multiple
No imputation

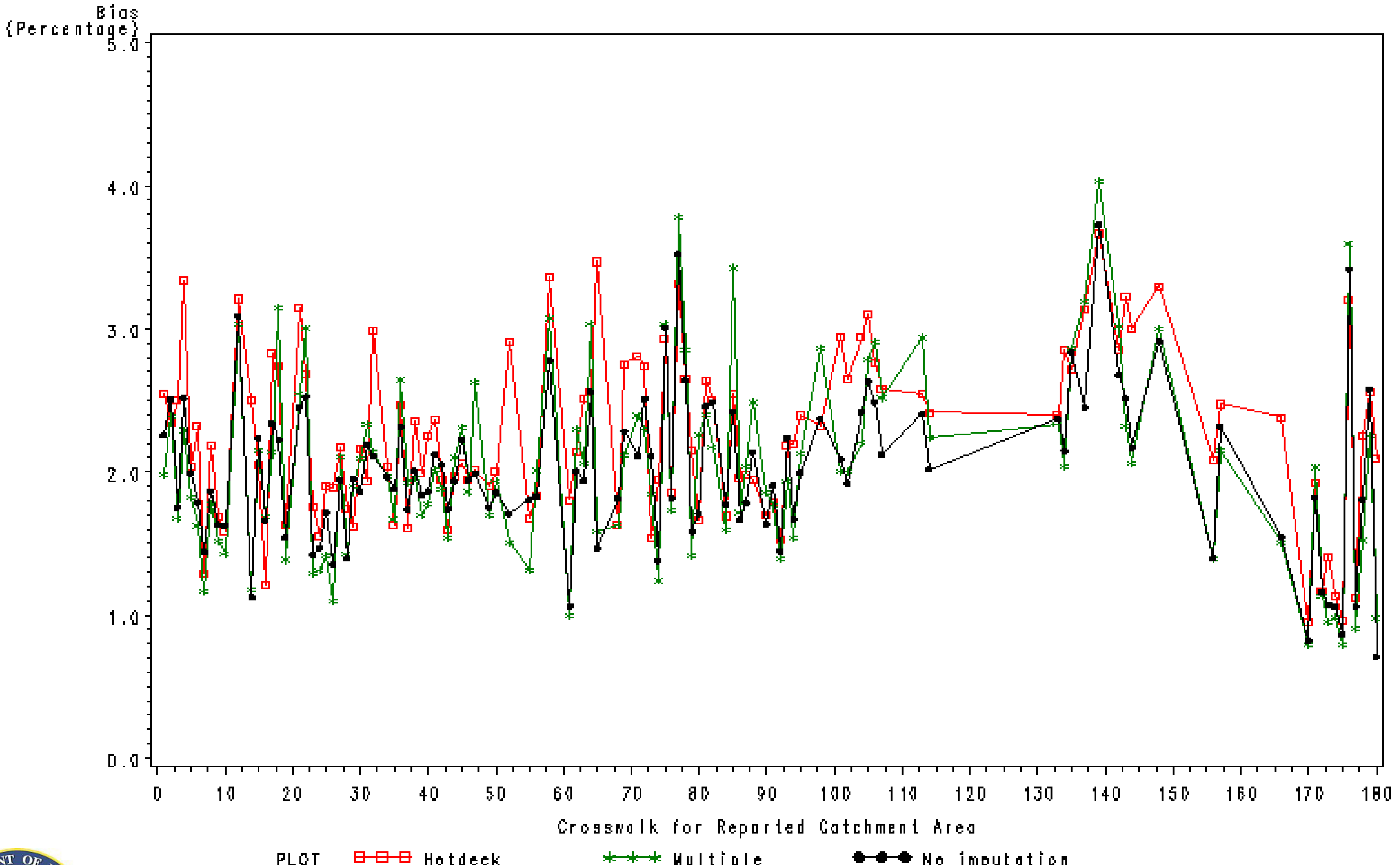
Mean of Diff for NMAR(Not Missing at Random) by XWCATCH



MATHEMATICA
Policy Research, Inc.

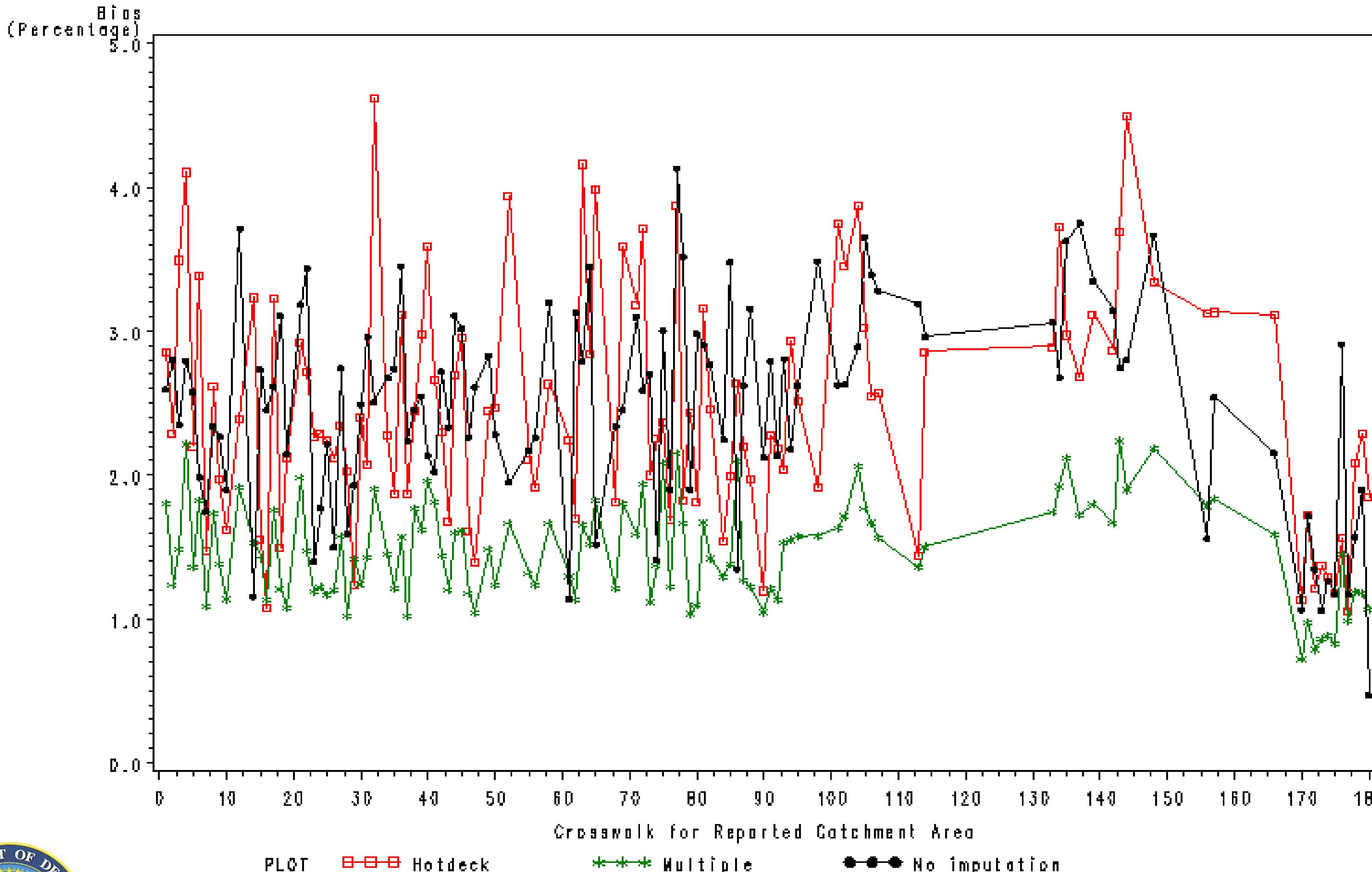
Bias for MAR(Missing at Random) by XWCATCH

Variable HO3023(In 1st yr:get care as soon as wanted)
with response rate of 88.7%



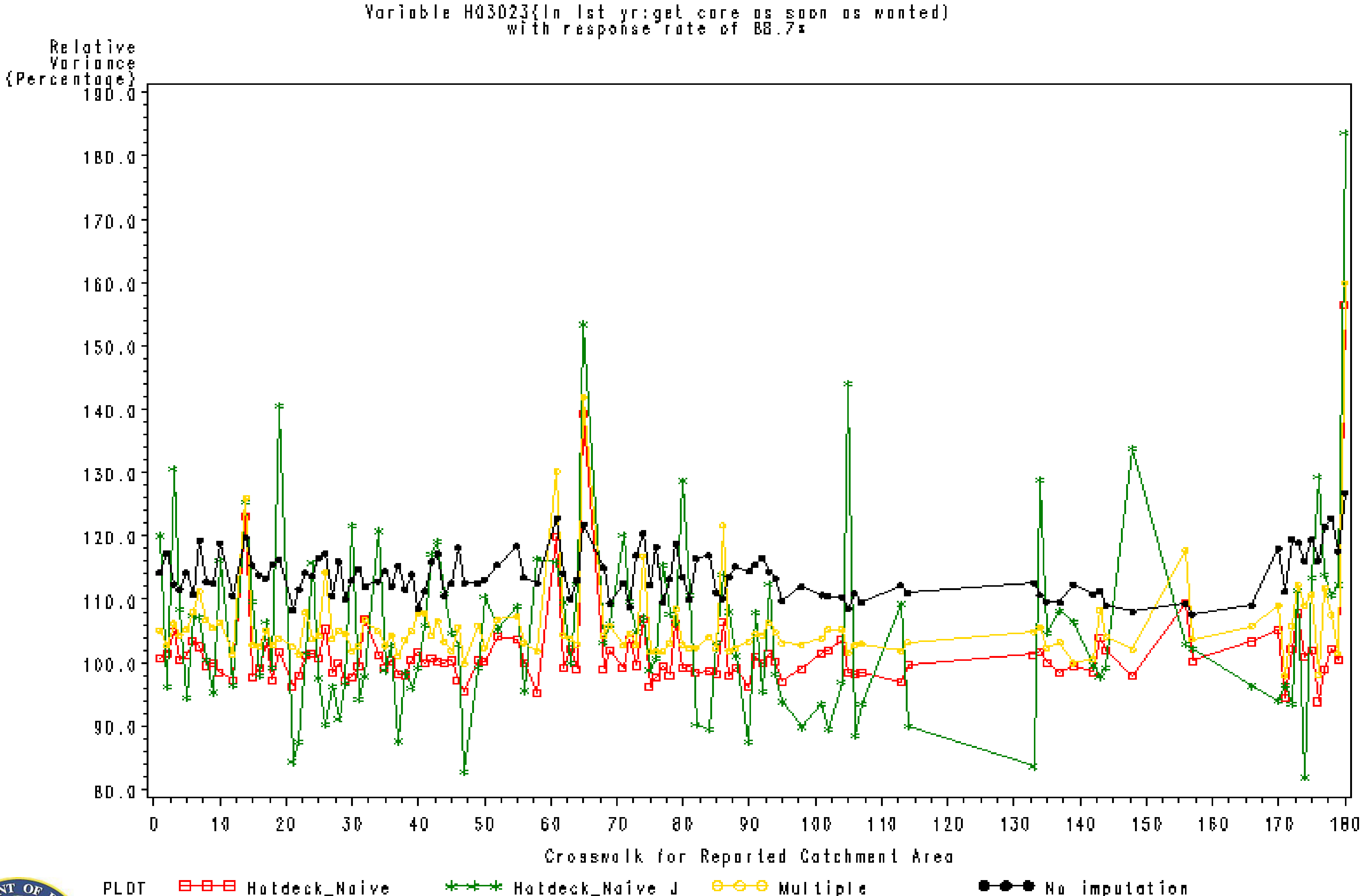
Bias for NMAR(Not Missing at Random) by XWCATCH

Variable H03023(In 1st yr:get care as soon as wanted)
with response rate of 88.7%

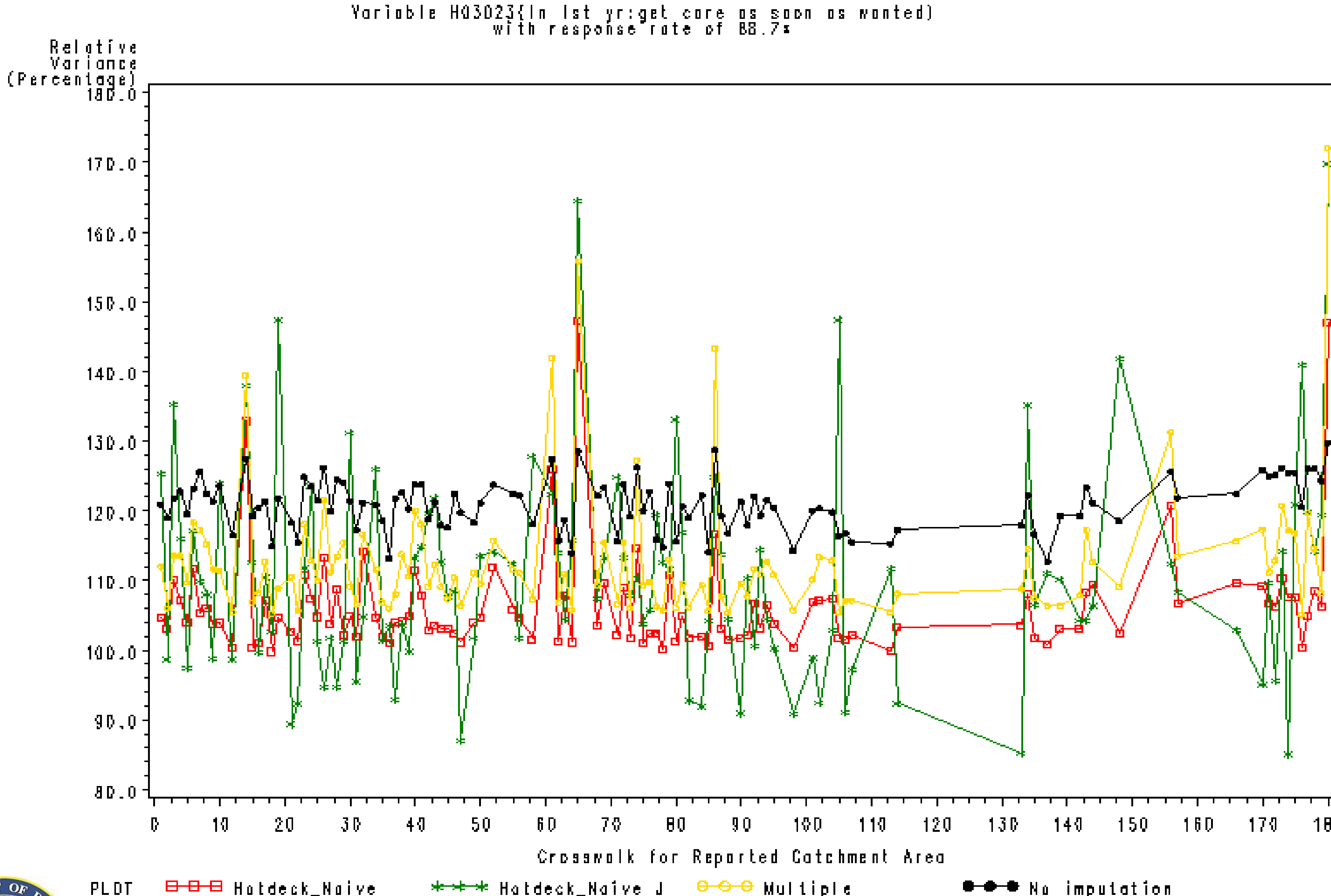


MATHEMATICA
Policy Research, Inc.

Relative Variance for MAR(Missing at Random) by XWCATCH



Relative Variance for NMAR(Not Missing at Random) by XWCATCH



MATHEMATICA
Policy Research, Inc.

Results

- Bias due to item missing
  - Not a major concern for univariate analyses in HCSDB
- Imputation increases effective sample sizes
  - Better precision
  - Need to account for imputation variability
- Multiple imputation
  - Good for bias reduction
  - Reasonable variance estimation
  - Multiple datasets

Results

- Hotdeck imputation
  - Produce a single survey data with complete rectangular data units
  - Provides consistent results
  - Variance estimation complicate: not user friendly
  - Alternatives: GVF or design effect type inflation factor

MATHEMATICA
Policy Research, Inc.