Evaluating Imputation Methods using ImpACT

First Case Study

UNECE Workshop on
Statistical Data Editing

August 31, 2020

Delivering insight through data for a better Canada

Darren Gray

Statistics Canada

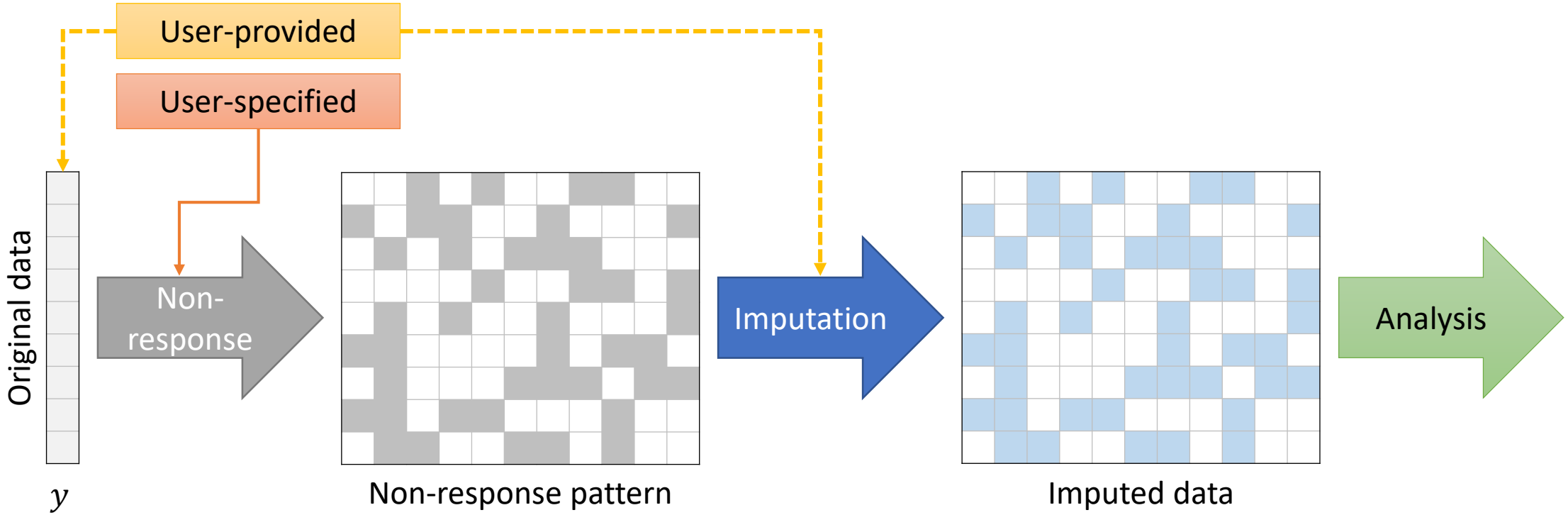
Introduction

- ImpACT: the Imputation Assessment and Comparison Tool
- Builds on similar work at Statistics Canada by Haziza and Stelmack
- Underlying framework presented at JSM (2019) using synthetic data
- Today's presentation covers first case study using survey data

ImpACT

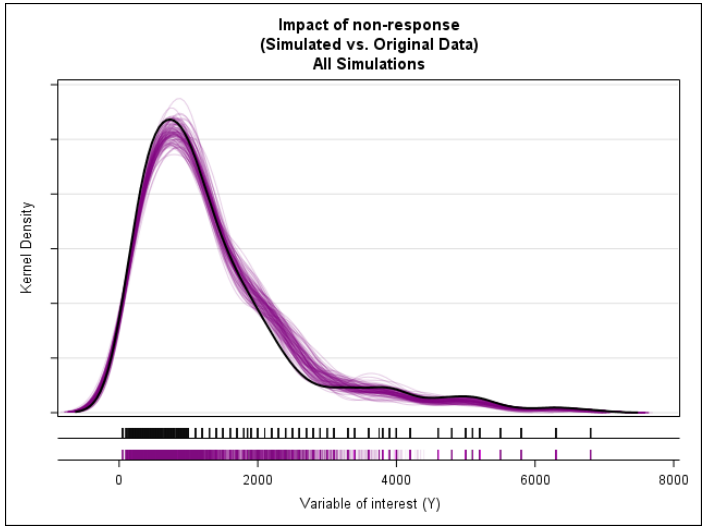
- Objective: provide survey methodologists with an easy-to-use, generalized tool to test and compare imputation methods in a controlled, simulation environment
- Consists of three modules: non-response, imputation and analysis
- Combining non-response patterns with specific analysis types can lead to a variety of useful insights

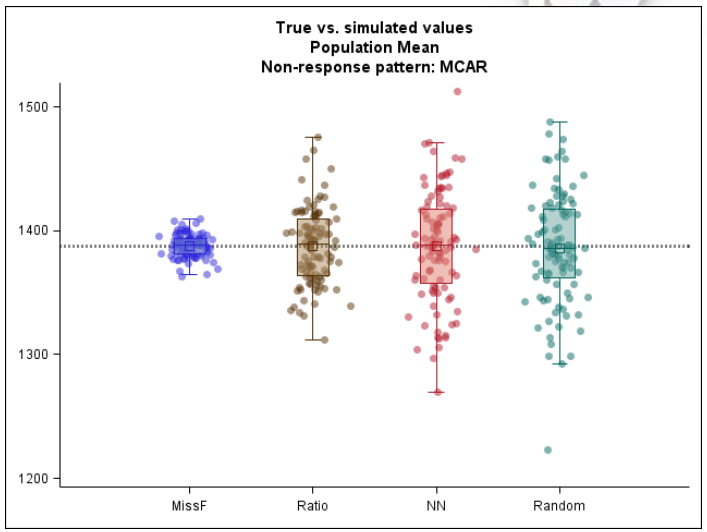
ImpACT Framework

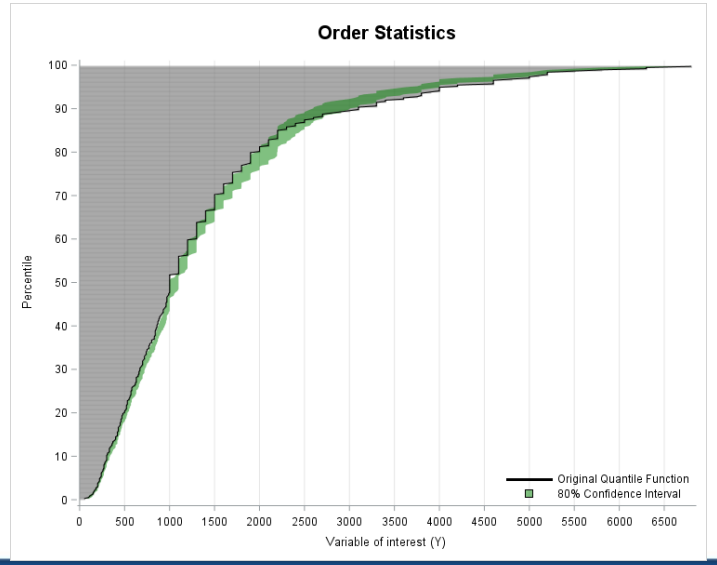


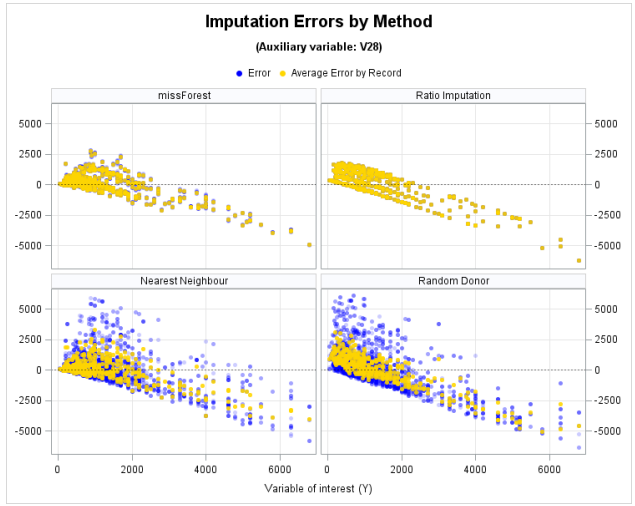
Three types of analysis:

- Distributional accuracy
- Estimation accuracy
- Prediction accuracy



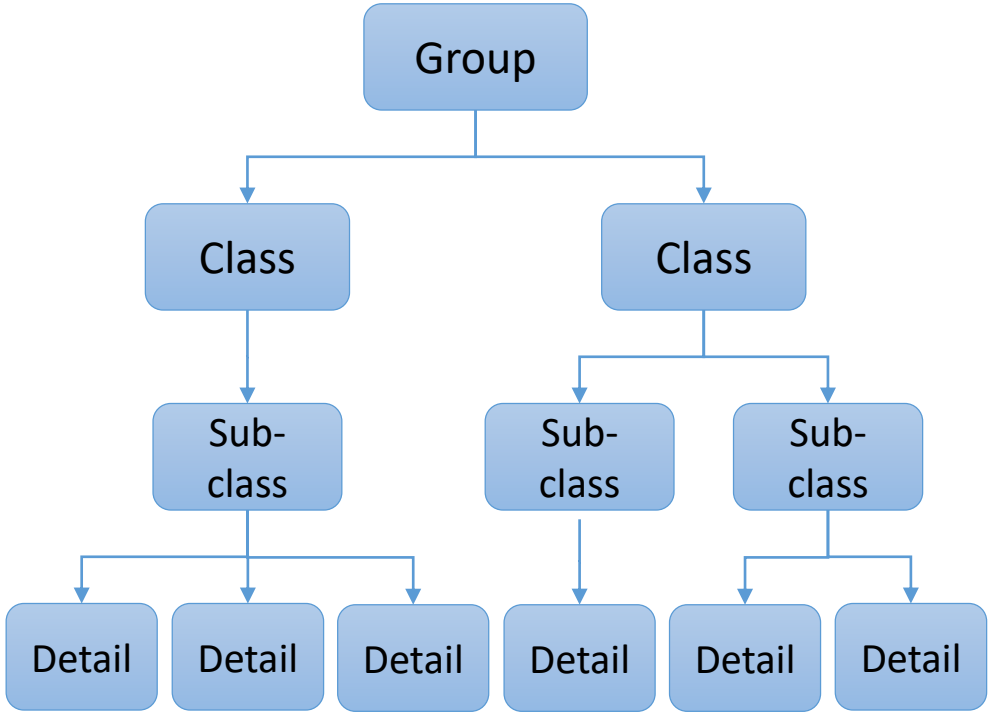






Retail Commodity Survey (RCS)

- Monthly survey measuring commodity sales
- Estimates broken down by commodity using North American Product Classification System (NAPCS)
- For each record, commodity sales obey additive structure



Retail Commodity Survey (RCS)

- Multi-stage data editing approach
- Missing values not imputed in one stage pass on to subsequent stages
- Altered imputation strategy for simulation purposes
- Study plan included a **general assessment** and **sub-method comparison**

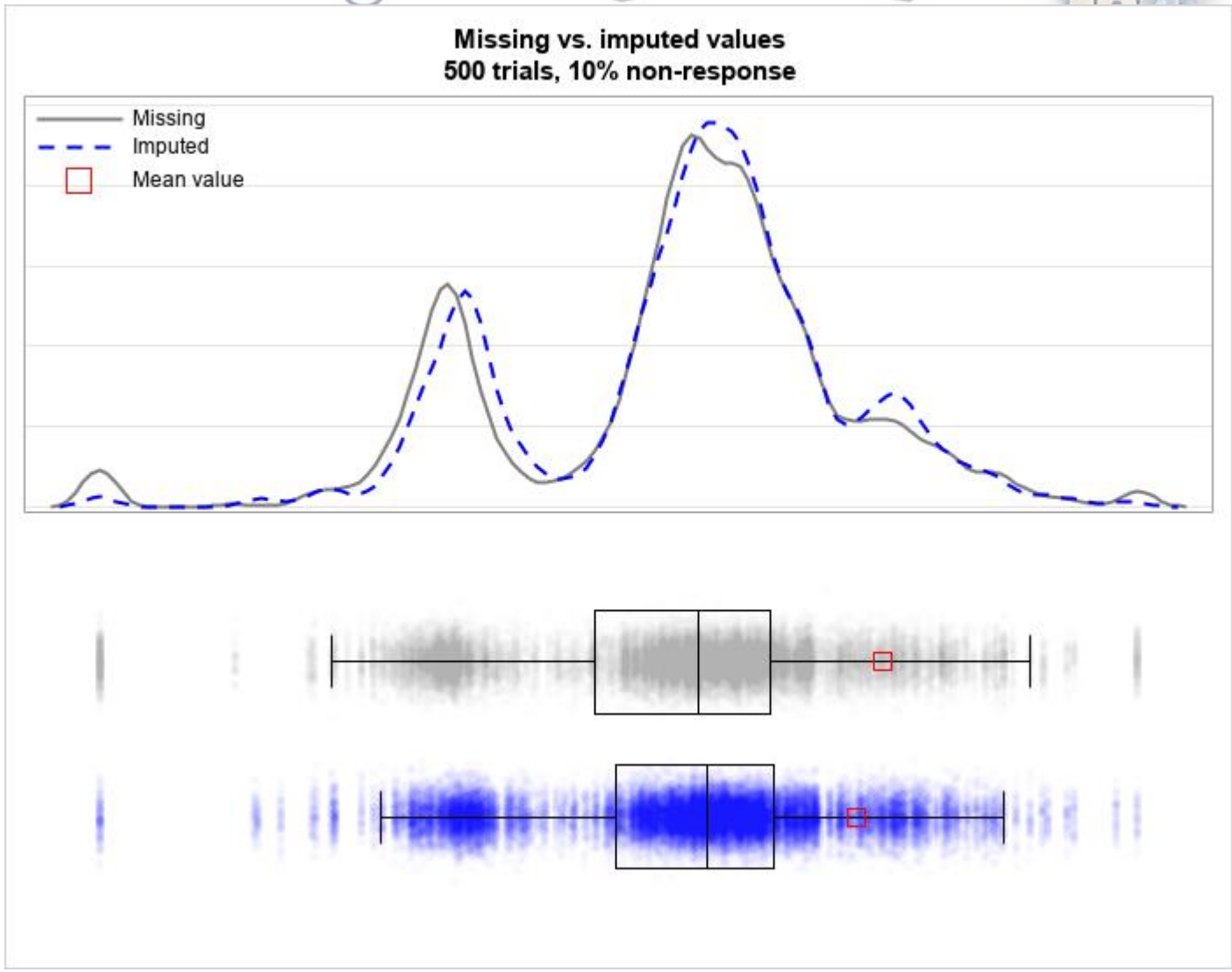
- Deductive imputation
- S1 • Historical imputation (previous year)
- S2 • Historical imputation (previous month)
- S3 • Donor imputation (modified)
- S4 • Ratio imputation (first set of imputation classes)
- S5 • Ratio imputation (second set of imputation classes)
- S6 • Percentage imputation
- Pro-rating

First study: general assessment

- **Goal**: Assess performance of overall strategy under simulated missing-completely-at-random (MCAR) non-response
- Three simulations (500 trials each) at non-response rates of 10%, 30% and 50%
- Within each trial, fixed non-response size and coordinated to ensure records were imputed the same number of times
- Analysis types presented: **distributional** and **estimation** accuracy

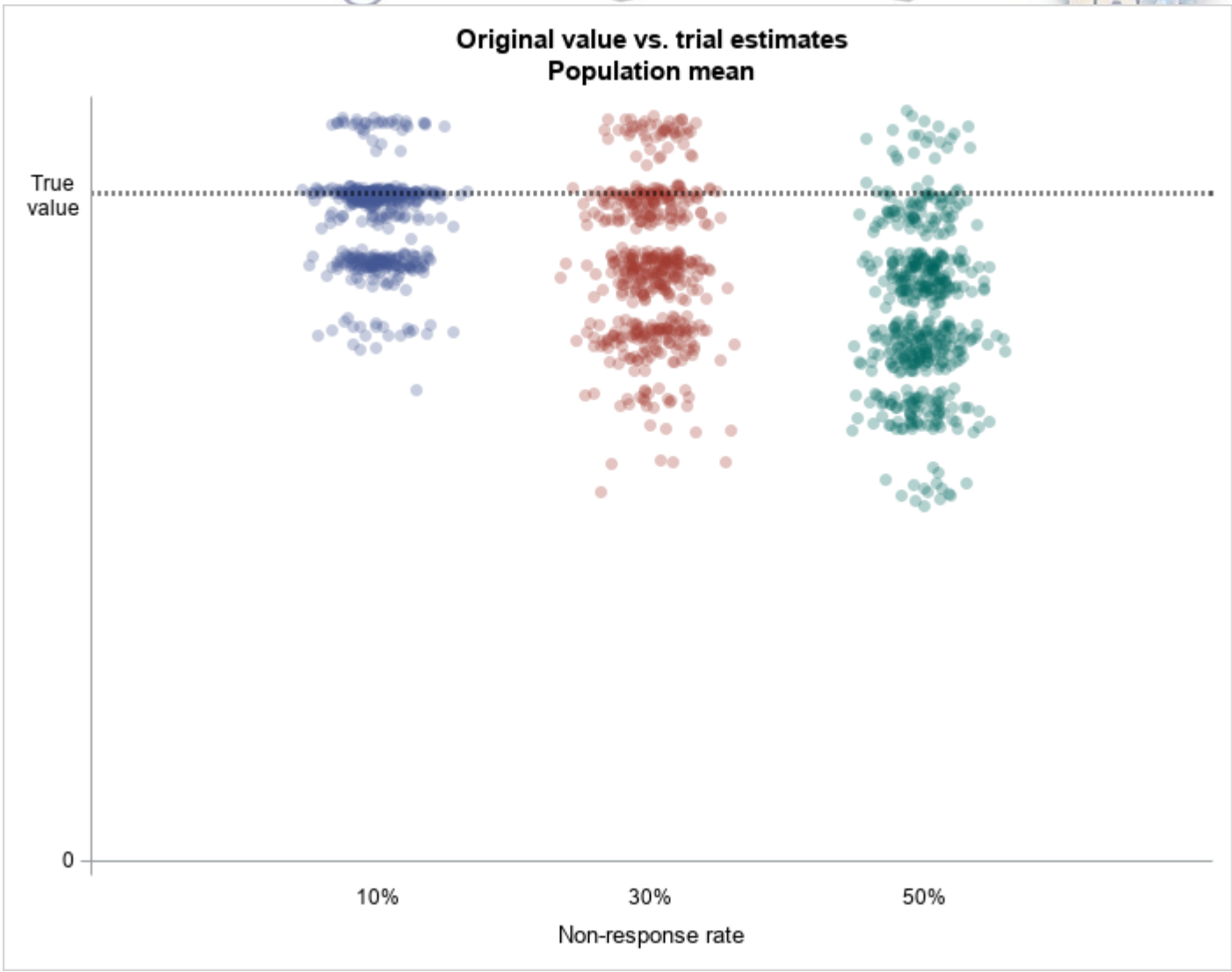
Distributional accuracy

- Compare distribution of missing values to imputed values, aggregated over all trials
- Options include kernel distribution, jitter plot, boxplot and mean comparison
- Combination of different elements gives a comprehensive picture
- (Note: values displayed on log scale)



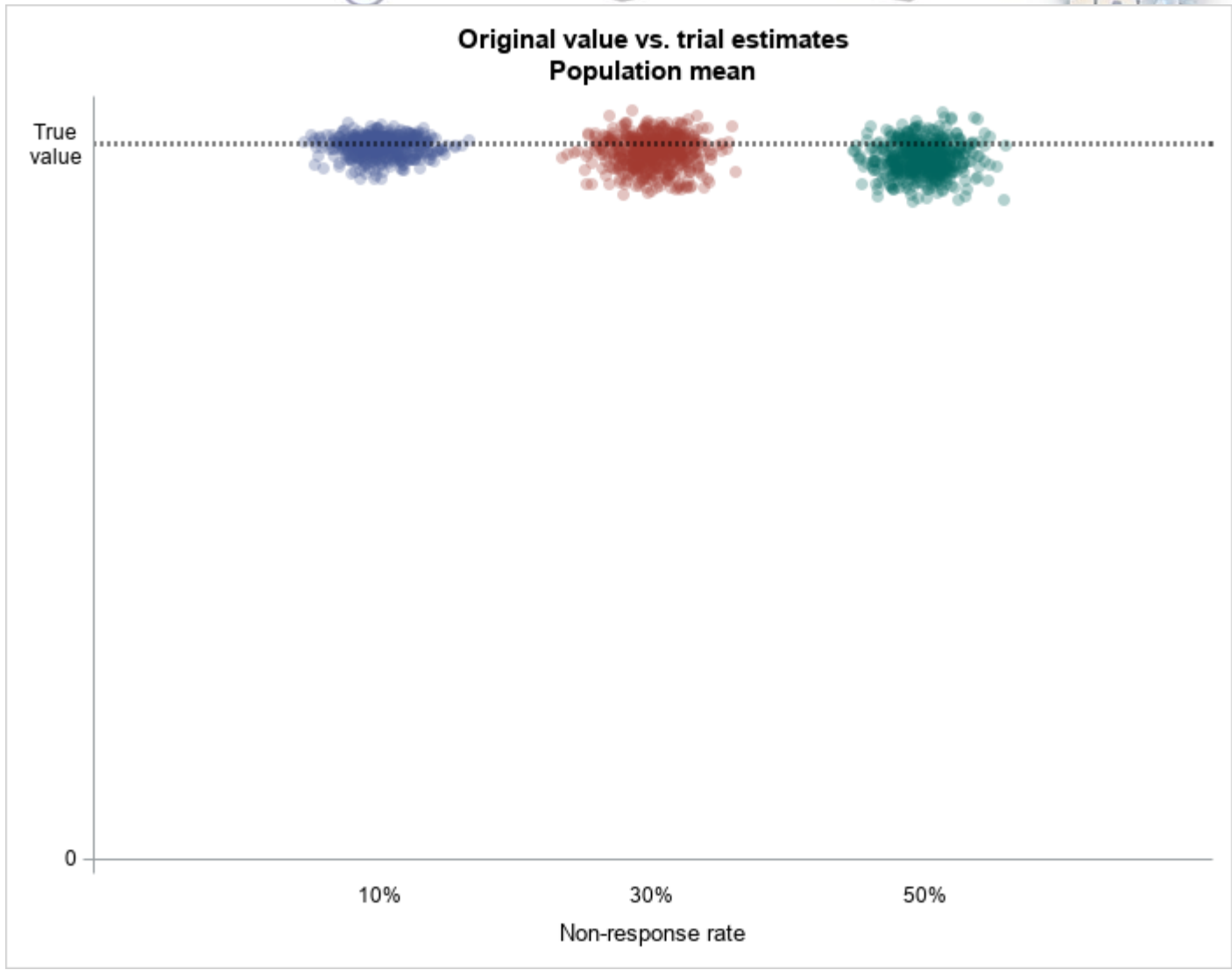
Estimation accuracy

- Calculate mean from original values
- Compare to mean generated by each trial
- Results show negative bias that increases in magnitude as non-response increases
- Investigation of clustering showed sensitivity to outliers



Estimation accuracy

- Removed outliers, repeated simulation
- Drastically reduced bias and variance



First study: inferential limitations

- Imputation rate by imputation stage showed that the simulations did not accurately reflect the true imputation process
- This can indicate a non-representative training set, or inaccurate non-response mechanism
- Rate of missingness for historical data differed between training set and the actual target of imputation

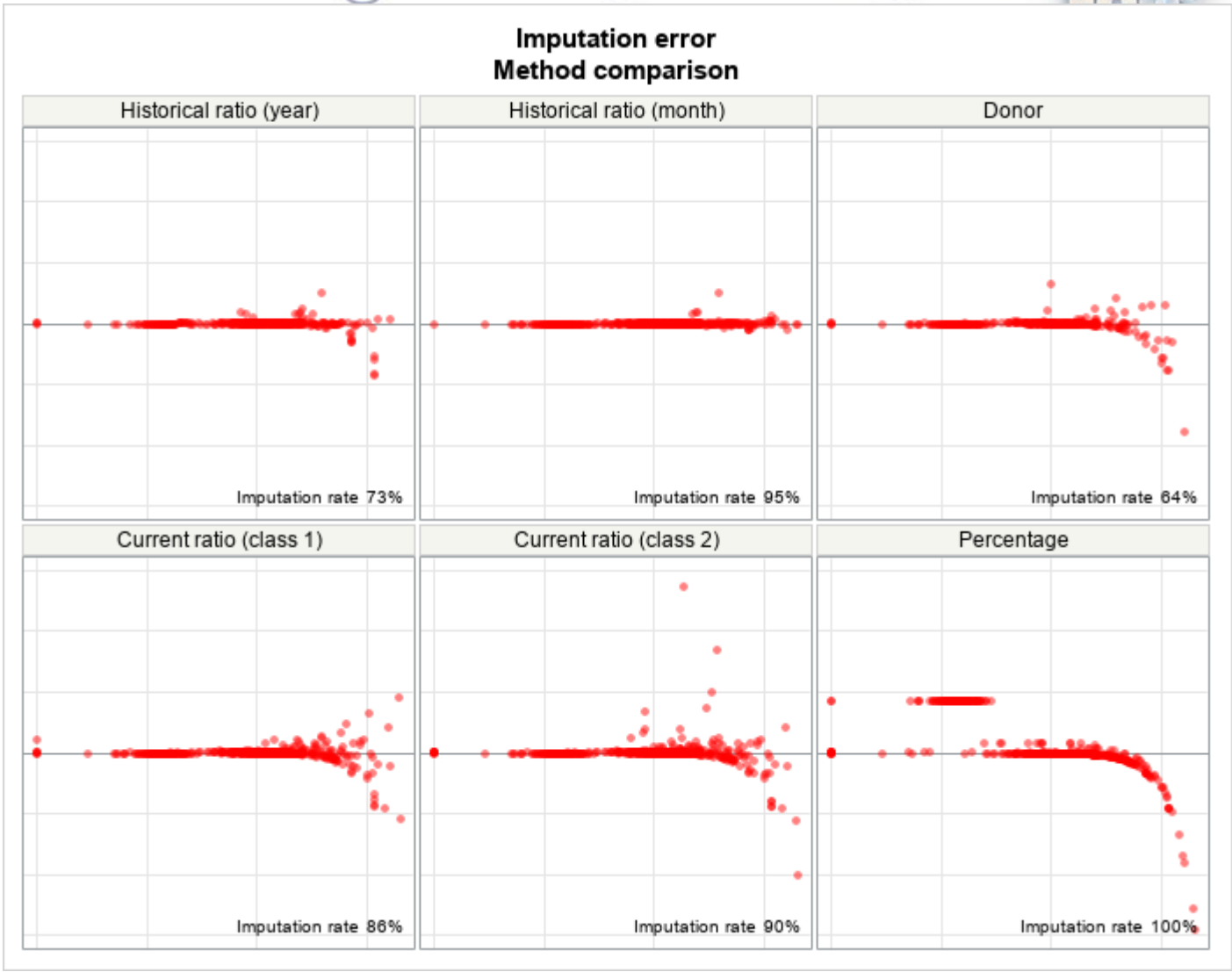
Stage	Simulation non-response rate			Actual
	10%	30%	50%	
S1	72.7	72.7	72.7	16.4
S2	25.5	25.5	25.5	3.0
S3	0.6	0.4	0.1	32.0
S4	0.4	0.6	0.9	31.6
S5	0.6	0.6	0.6	5.7
S6	0.3	0.3	0.3	11.3

Second study: sub-method comparison

- **Goal**: Investigate sub-methods using a leave-one-out cross-validation approach
- One trial for each record
- Analysis type presented: **predictive accuracy**
- Challenge: comparing methods that only impute a subset of all records

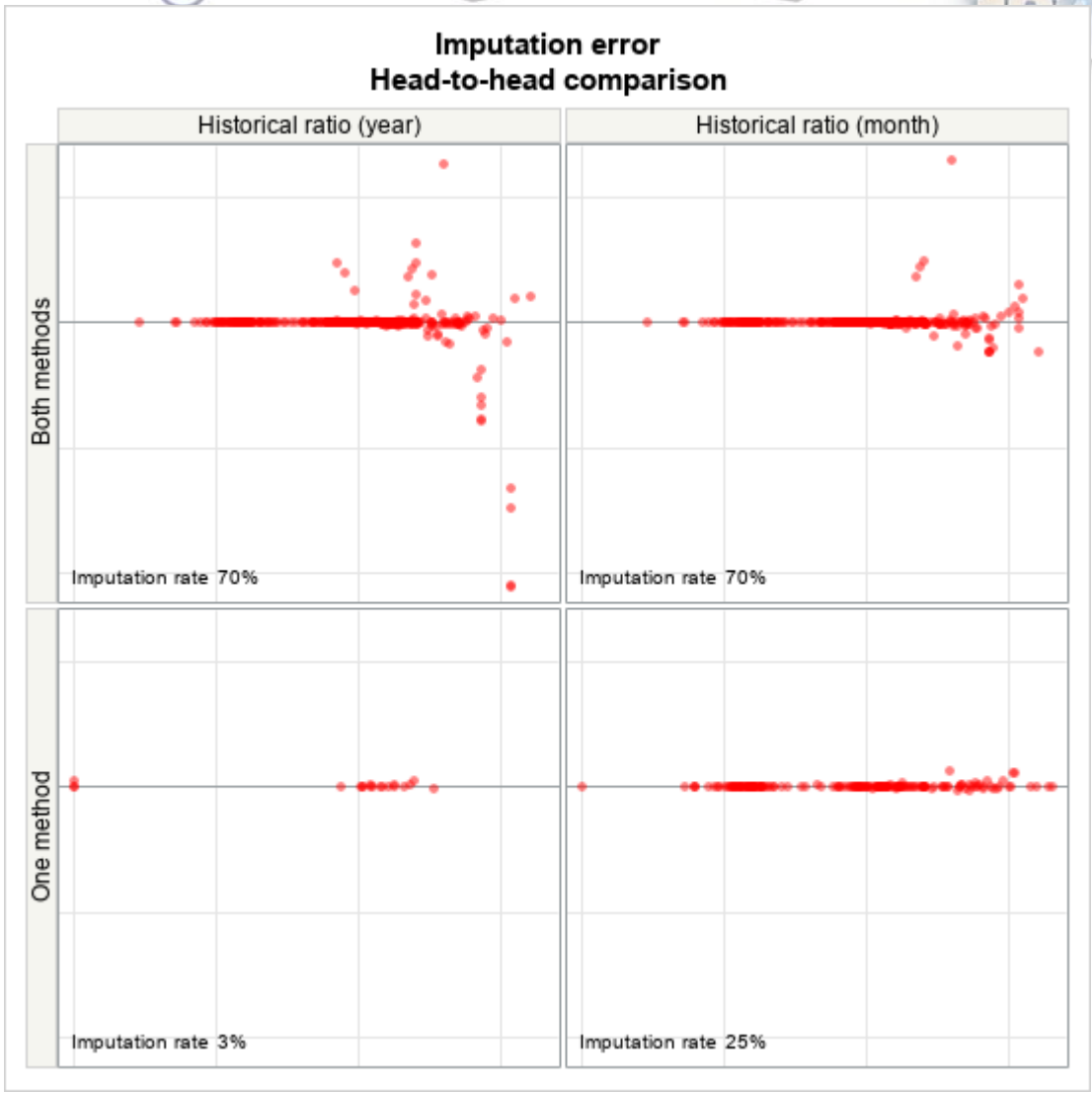
Predictive accuracy

- Compare imputation errors against original values (log scale) across all methods
- Imputation rates vary by method
- Only stage raising concerns is the final one (percentage imputation)
- Should we change the order?



Predictive accuracy

- Comparison of historical ratio methods
- Use of previous month data outperforms previous year (could be specific to our reference period)



Predictive accuracy

- Compare donor imputation to first run of ratio imputation
- Ratio imputation method outperforms donor when limited to records imputed by both methods
- Donor imputation has other benefits

Conclusions

- ImpACT can provide insights into behaviour of survey imputation process
  - Insights into accuracy of imputation strategy
  - Lead to further investigation
  - Users should be aware of inferential limitations

- Challenges posed by survey data and complex imputation design:
  - Missingness in auxiliary variables
  - Multi-stage imputation strategy required new assessment tools
  - Multivariate sub-method (donor imputation) required modification

THANK YOU!

Contact:
darren.gray@canada.ca

The content of this presentation represents the position of the author and may not necessarily represent that of Statistics Canada.