Calibrating Big Data for Population Inference: Applying Quasi-randomization Approach to Naturalistic Driving Data using Bayesian Additive Regression Trees

Ali Rafei[1] Carol A. C. Flannagan[3] Michael R. Elliott[2]

[1]Program in Survey Methodology, University of Michigan

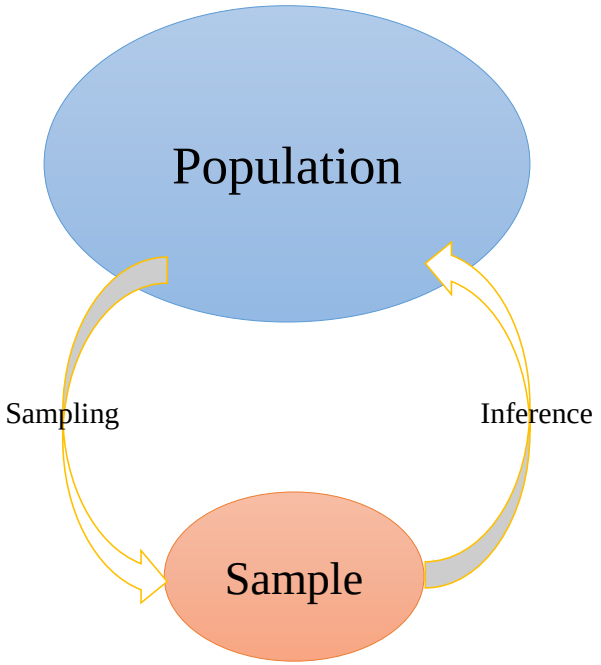
[2]Department of Biostatistics, University of Michigan

[3]University of Michigan Transportation Research Institute

BigSurv Meeting
October 2018

Big Data vs survey data

- Probability sampling is the "gold standard" for population inference (Neyman, 1934).
  1. The upward trends of non-response rate
  2. The rising cost and complexity



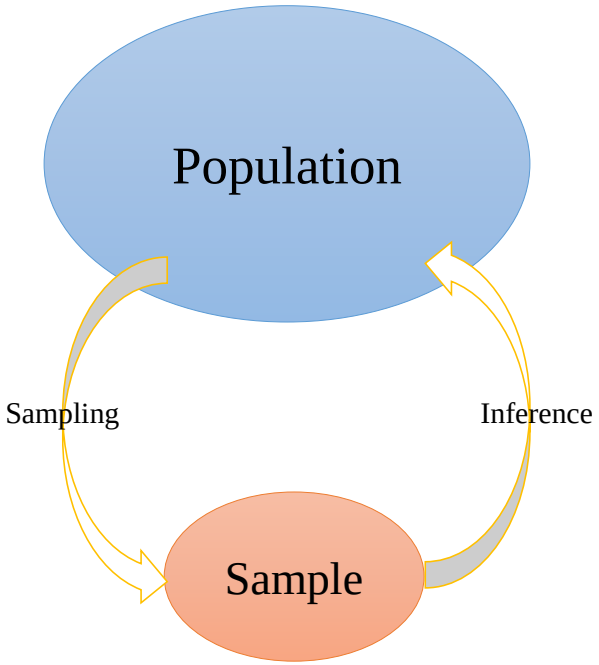
- Methods are developed to adjust for selection bias, where the "gold standard" is violated.
  1. Post-survey adjustment in probability samples e.g. GREG
  2. Calibration methods for nonprobability samples e.g. pseudo-weighting

- Parallel to surveys, Big Data are increasingly available e.g. social media, sensor data, transaction records
  1. Predictive analysis and small area estimation
  2. "Big challenge" for population inference

Big Data vs survey data

- Probability sampling is the "gold standard" for population inference (Neyman, 1934).
  1. The upward trends of non-response rate
  2. The rising cost and complexity

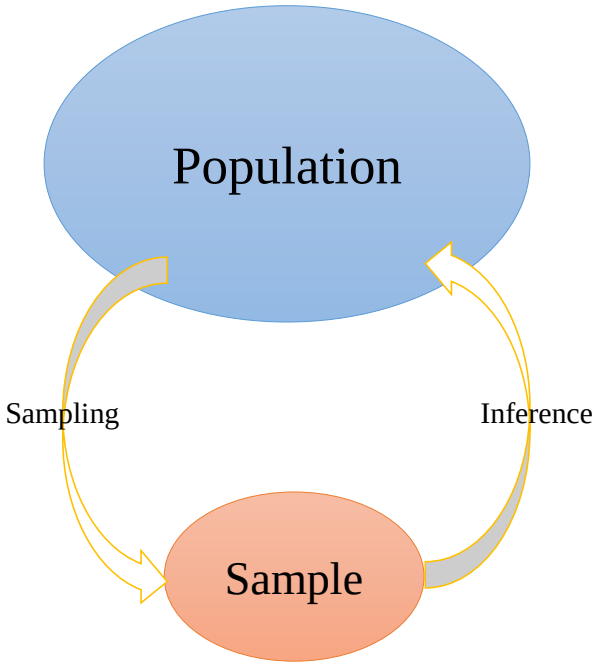
- Methods are developed to adjust for selection bias, where the "gold standard" is violated.
  1. Post-survey adjustment in probability samples e.g. GREG
  2. Calibration methods for nonprobability samples e.g. pseudo-weighting



- Parallel to surveys, Big Data are increasingly available e.g. social media, sensor data, transaction records
  1. Predictive analysis and small area estimation
  2. "Big challenge" for population inference

Big Data vs survey data

- Probability sampling is the "gold standard" for population inference (Neyman, 1934).
  1. The upward trends of non-response rate
  2. The rising cost and complexity

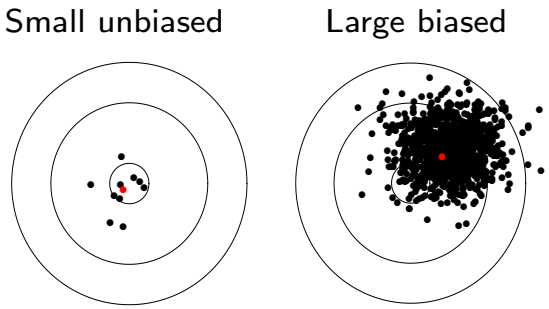


- Methods are developed to adjust for selection bias, where the "gold standard" is violated.
  1. Post-survey adjustment in probability samples e.g. GREG
  2. Calibration methods for nonprobability samples e.g. pseudo-weighting

- Parallel to surveys, Big Data are increasingly available e.g. social media, sensor data, transaction records
  1. Predictive analysis and small area estimation
  2. "Big challenge" for population inference

Big Data and finite population inference

- **Selection bias** is a major concern.
  1. Lack of a known random selection mechanism
  2. Lack of control over data collection process

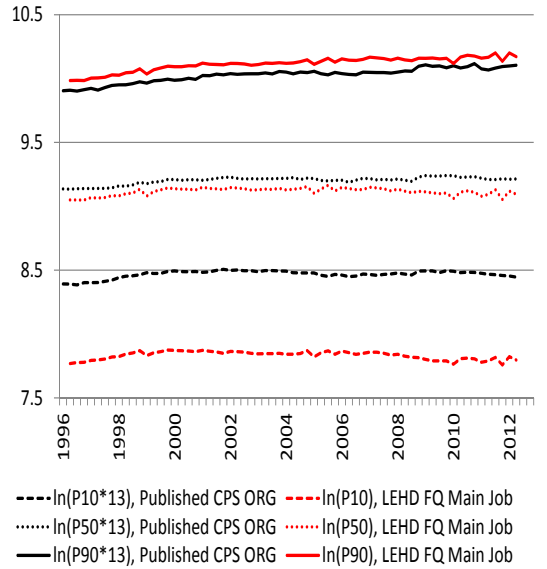


- Does bigger necessarily imply better?
  1. Literary Digest poll (n=2.3M) vs Gallup poll (n=3K)
  2. CPS (c.r.=0.05%) vs LEHD (c.r.=90%)

- **Big Data Paradox**: "The bigger the data, the more certain we will miss our target" (Xiao-Li Meng).

- **Objective**: To improve the representativeness of naturalistic driving data in SPMD based on the NHTS as benchmark using quasi-randomization approach.

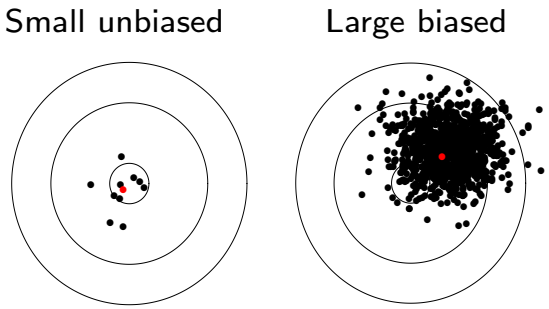
CPS vs LEHD: earnings percentiles



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

Big Data and finite population inference

- **Selection bias** is a major concern.
  1. Lack of a known random selection mechanism
  2. Lack of control over data collection process

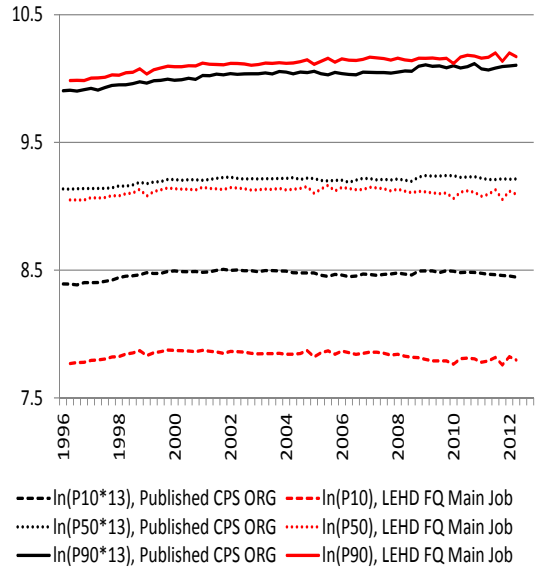


- Does bigger necessarily imply better?
  1. Literary Digest poll (n=2.3M) vs Gallup poll (n=3K)
  2. CPS (c.r.=0.05%) vs LEHD (c.r.=90%)

- **Big Data Paradox**: "The bigger the data, the more certain we will miss our target" (Xiao-Li Meng).

- **Objective**: To improve the representativeness of naturalistic driving data in SPMD based on the NHTS as benchmark using quasi-randomization approach.

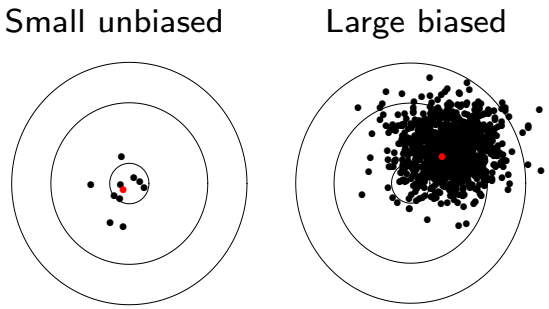
CPS vs LEHD: earnings percentiles



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

Big Data and finite population inference

- **Selection bias** is a major concern.
  1. Lack of a known random selection mechanism
  2. Lack of control over data collection process

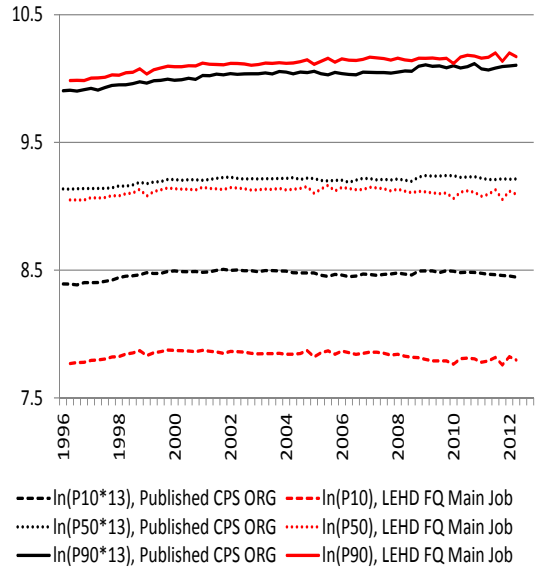


- Does bigger necessarily imply better?
  1. Literary Digest poll (n=2.3M) vs Gallup poll (n=3K)
  2. CPS (c.r.=0.05%) vs LEHD (c.r.=90%)

- **Big Data Paradox**: "The bigger the data, the more certain we will miss our target" (Xiao-Li Meng).

- **Objective**: To improve the representativeness of naturalistic driving data in SPMD based on the NHTS as benchmark using quasi-randomization approach.

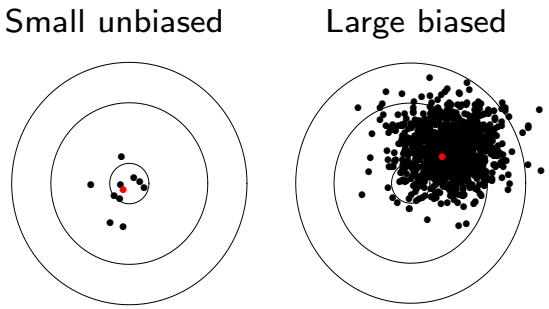
CPS vs LEHD: earnings percentiles



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

Big Data and finite population inference

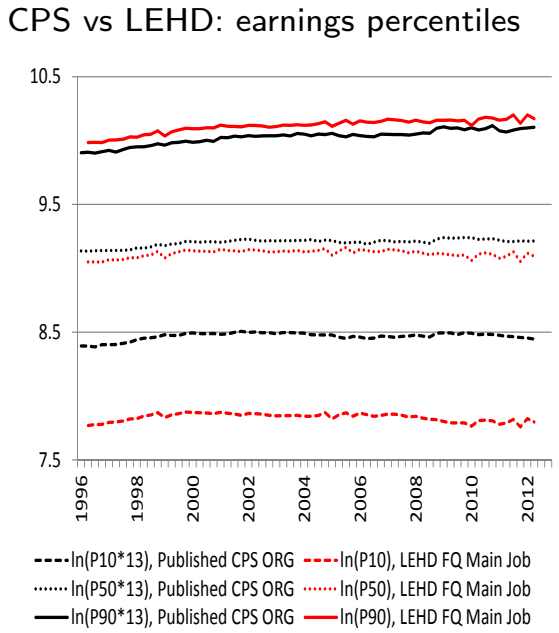
- **Selection bias** is a major concern.
  1. Lack of a known random selection mechanism
  2. Lack of control over data collection process



- Does bigger necessarily imply better?
  1. Literary Digest poll (n=2.3M) vs Gallup poll (n=3K)
  2. CPS (c.r.=0.05%) vs LEHD (c.r.=90%)

- **Big Data Paradox**: "The bigger the data, the more certain we will miss our target" (Xiao-Li Meng).

- **Objective**: To improve the representativeness of naturalistic driving data in SPMD based on the NHTS as benchmark using quasi-randomization approach.



Naturalistic Driving Studies (NDS)

- Continuously monitoring of driving behaviors by instrumented vehicles
- A rich resource for testing novel transportation technologies, and exploring crash causality, traffic safety, and travel dynamics.
- Launched in Sep 2012 by UMTRI, SPMD is among the largest NDS, encompassing over 3,000 instrumented vehicles.
- SPMD characterizes real-world implementation of automated and connected vehicles, with the primary aim of testing DSRC-based connected vehicle communication technology.
- $\sim$6 million trips were captured within average participation time of $\sim$1 person-year.
- Participants were mostly volunteers from Ann Arbor, selected through a combination of convenience and snowball sampling methods.

Naturalistic Driving Studies (NDS)

- Continuously monitoring of driving behaviors by instrumented vehicles
- A rich resource for testing novel transportation technologies, and exploring crash causality, traffic safety, and travel dynamics.
- Launched in Sep 2012 by UMTRI, SPMD is among the largest NDS, encompassing over 3,000 instrumented vehicles.
- SPMD characterizes real-world implementation of automated and connected vehicles, with the primary aim of testing DSRC-based connected vehicle communication technology.
- $\sim$6 million trips were captured within average participation time of $\sim$1 person-year.
- Participants were mostly volunteers from Ann Arbor, selected through a combination of convenience and snowball sampling methods.

Naturalistic Driving Studies (NDS)

- Continuously monitoring of driving behaviors by instrumented vehicles
- A rich resource for testing novel transportation technologies, and exploring crash causality, traffic safety, and travel dynamics.
- Launched in Sep 2012 by UMTRI, SPMD is among the largest NDS, encompassing over 3,000 instrumented vehicles.
- SPMD characterizes real-world implementation of automated and connected vehicles, with the primary aim of testing DSRC-based connected vehicle communication technology.
- $\sim$6 million trips were captured within average participation time of $\sim$1 person-year.
- Participants were mostly volunteers from Ann Arbor, selected through a combination of convenience and snowball sampling methods.

Review of existing approaches

- Elliott, M., and Valliant, R. (2017) provides an extensive review of methods used for calibrating non-probability samples.
- There are two general approaches:
  1. **Quasi-randomization:**
     The inclusion probabilities are modeled and predicted to construct pseudo-weights.
     - Pseudo-weighting
     - Predictive Mean Matching (PMM)
  2. **Super-population:**
     The analytic variable is modeled to be predicted for nonsample units.

Review of existing approaches

- Elliott, M., and Valliant, R. (2017) provides an extensive review of methods used for calibrating non-probability samples.
- There are two general approaches:
  1. **Quasi-randomization:**
     The inclusion probabilities are modeled and predicted to construct pseudo-weights.
     - Pseudo-weighting
     - Predictive Mean Matching (PMM)
  2. **Super-population:**
     The analytic variable is modeled to be predicted for nonsample units.

Review of existing approaches

- Elliott, M., and Valliant, R. (2017) provides an extensive review of methods used for calibrating non-probability samples.
- There are two general approaches:
  1. **Quasi-randomization:**
     The inclusion probabilities are modeled and predicted to construct pseudo-weights.
     - Pseudo-weighting
     - Predictive Mean Matching (PMM)

Quasi-randomization

- Let S_i and S_i^* denote the indicator for probability sample and non-probability sample, respectively. Then, combining the probability and non-probability samples, we define Z_i to be the indicator for non-probability cases, given $S_i^* + S_i = 1$.

Pseudo-weighting:

$$\tilde{\omega}_i = 1/\hat{P}(S_i^* = 1|x_i = x_0) \propto \omega_i \times \frac{\hat{P}(Z_i = 0|x_i = x_0)}{\hat{P}(Z_i = 1|x_i = x_0)}$$

where $\omega_i = 1/P(S_i = 1|x_i = x_0)$.

Predictive mean matching:

Let $p_0, p_1, ..., p_{100}$ to be the percentiles of probabilities of selection in the probability sample,

$$\tilde{\omega}_i = \frac{q_k / \sum_k q_k}{n_k / n}, i \in k$$

where $q_k = \sum_j \omega_j I(p_{k-1} < \omega_j^{-1} \leq p_k)$.

Generalized Linear Regression

- Need to compute $\hat{P}(S_i|x_i = x_0)$ and $\hat{P}(Z_i|x_i = x_0)$.
- Can compute $\hat{P}(S_i|x_i = x_0)$ without error if x contains all design variables and sampling mechanism is known; otherwise use logistic or beta regression.
- Compute $\hat{P}(Z_i|x_i = x_0)$ using logistic regression
- For variance estimation, use a jackknife estimator across both the non-probability and probability sample
  - Non-probability sample: single stratum with IID observations.
  - Probability sample: Follow the sample design.

Bayesian Additive Regression Trees (BART)

- To minimize model misspecification when estimating the probabilities of selection and sample membership, we use BART (Chipman et al. 2007).
- Provides a strong flexible predictive tool by capturing non-linear relationships as well as high-order interaction effects.
- The idea is based on the sum-of-trees model.

BART approximates the outcome Y_i as:

$$Y_i = \sum_{j=1}^{m} f(x_i, T_j, M_j) + \epsilon_i \quad \textit{where} \quad \epsilon_i \sim N(0, \sigma^2)$$

- BART is a Bayesian approach that assigns a prior distribution to T, M, and σ.
- For binary outcomes, a *probit* link is typically used.

Variance estimation in pseudo-weighting

- To incorporate variability in both outcome and pseudo-weights, we employ a conditional variance approach:

$$Var(\hat{y}) = E\{Var(\hat{y}|w)\} + Var\{E(\hat{y}|w)\}$$

where the first part can be obtained through the TSL method:

$$\widehat{E}\{Var(\hat{y}|w)\} = var_{TSL}(\hat{y}_w)$$

and the second part is estimated based on the MCMC of posterior in BART:

$$\widehat{Var}\{E(\hat{y}|w)\} = \frac{1}{K-1}\sum_{j=1}^{K}(\hat{y}_{w_j} - \hat{y}_w)^2$$

$$w_{ij} = 1/\hat{P}_j(S_i = 1|x_i)\frac{\hat{P}_j(Z_i = 0|x_i)}{\hat{P}_j(Z_i = 1|x_i)}, \quad \forall j = 1, 2, ..., K$$

Construction of pseudo-weights

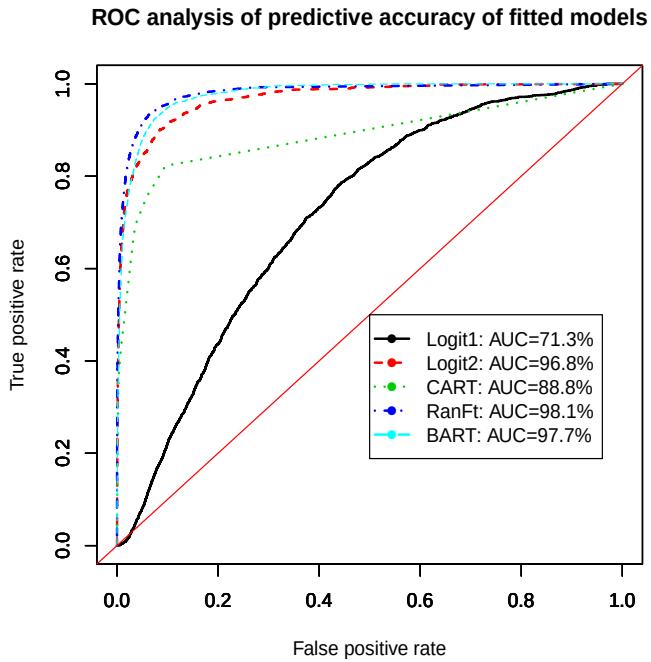
Comparing BART with alternative models in two steps of pseudo-weighting:

(1) $\hat{P}(S_i^* = 1 | x_i = x_0)$:

Model	RMSE	R^2
Original scale of response		
Linear Reg I	0.0190	5.95
Linear Reg II	0.0190	6.51
Poisson Reg I	0.0190	5.95
Poisson Reg II	0.0190	6.53
Beta Reg I	0.0191	5.02
Beta Reg II	0.0191	5.32
BART	0.0189	**7.43**
Log Scale of response		
Linear Reg I	1.350	14.99
Linear Reg II	1.346	15.56
BART	1.329	**17.66**

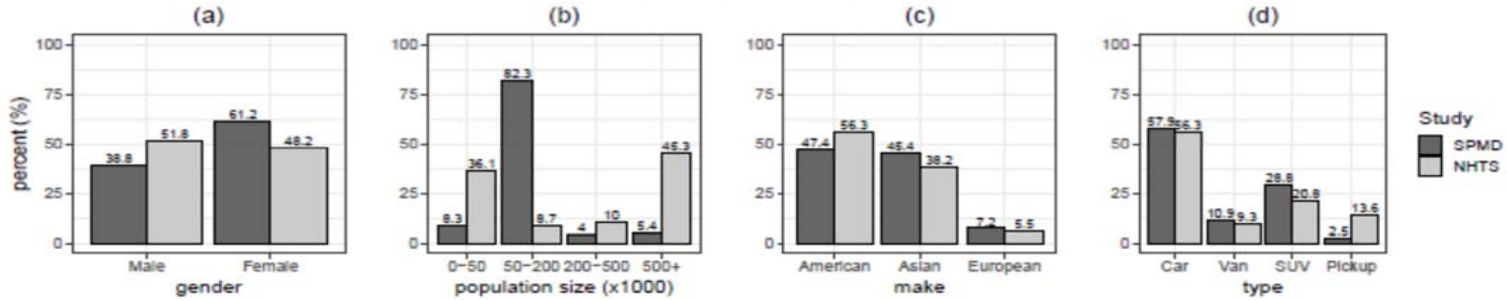
I: main effects in the model
II: two-way interaction effects were included

(2) $\hat{P}(Z_i = 1 | x_i = x_0)$:

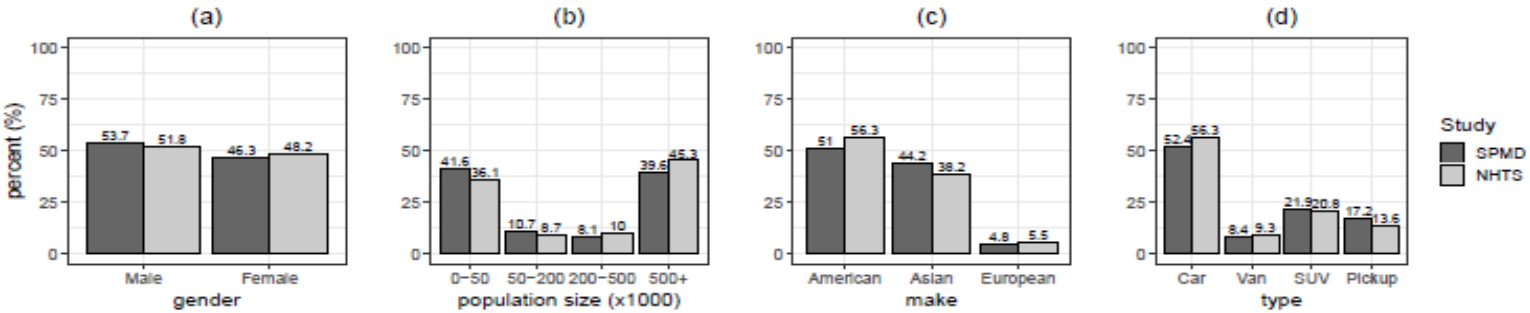


Distribution of common categorical auxiliary variables

Distribution of unweighted categorical covariates in SPMD vs NHTS:

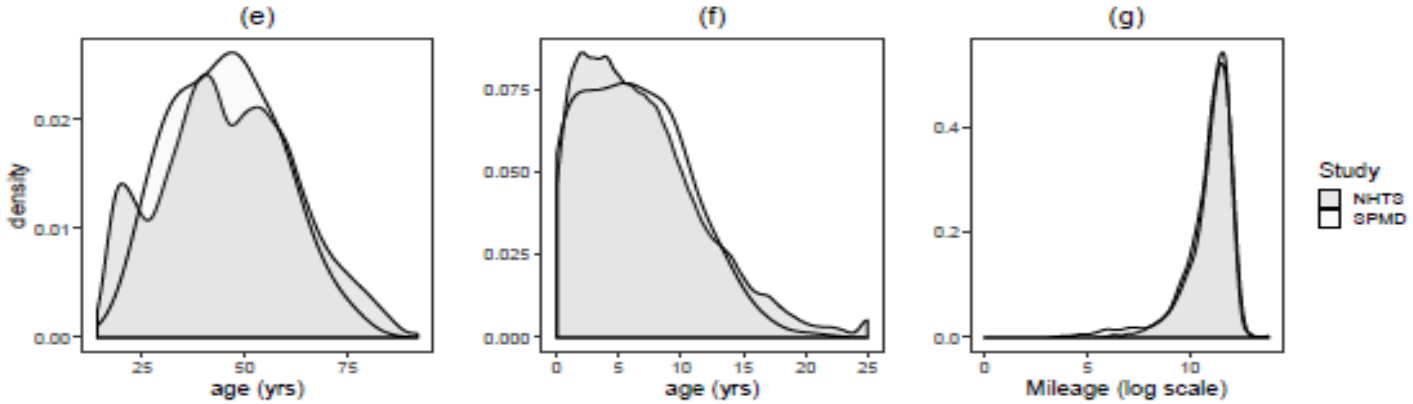


Distribution of categorical covariates using pseudo-weights:

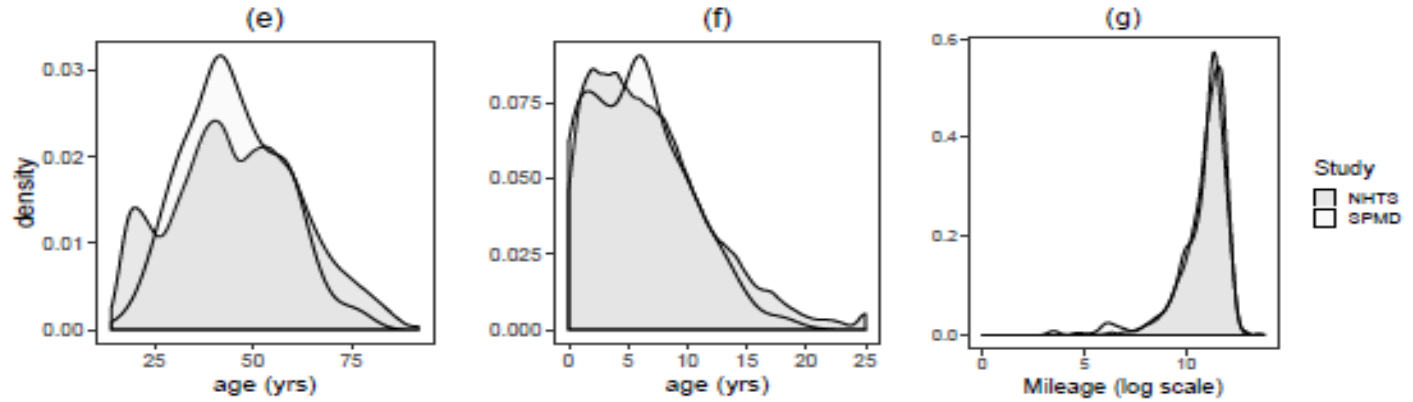


Distribution of common continuous auxiliary variables

Kernel density of unweighted continuous covariates in SPMD vs NHTS:

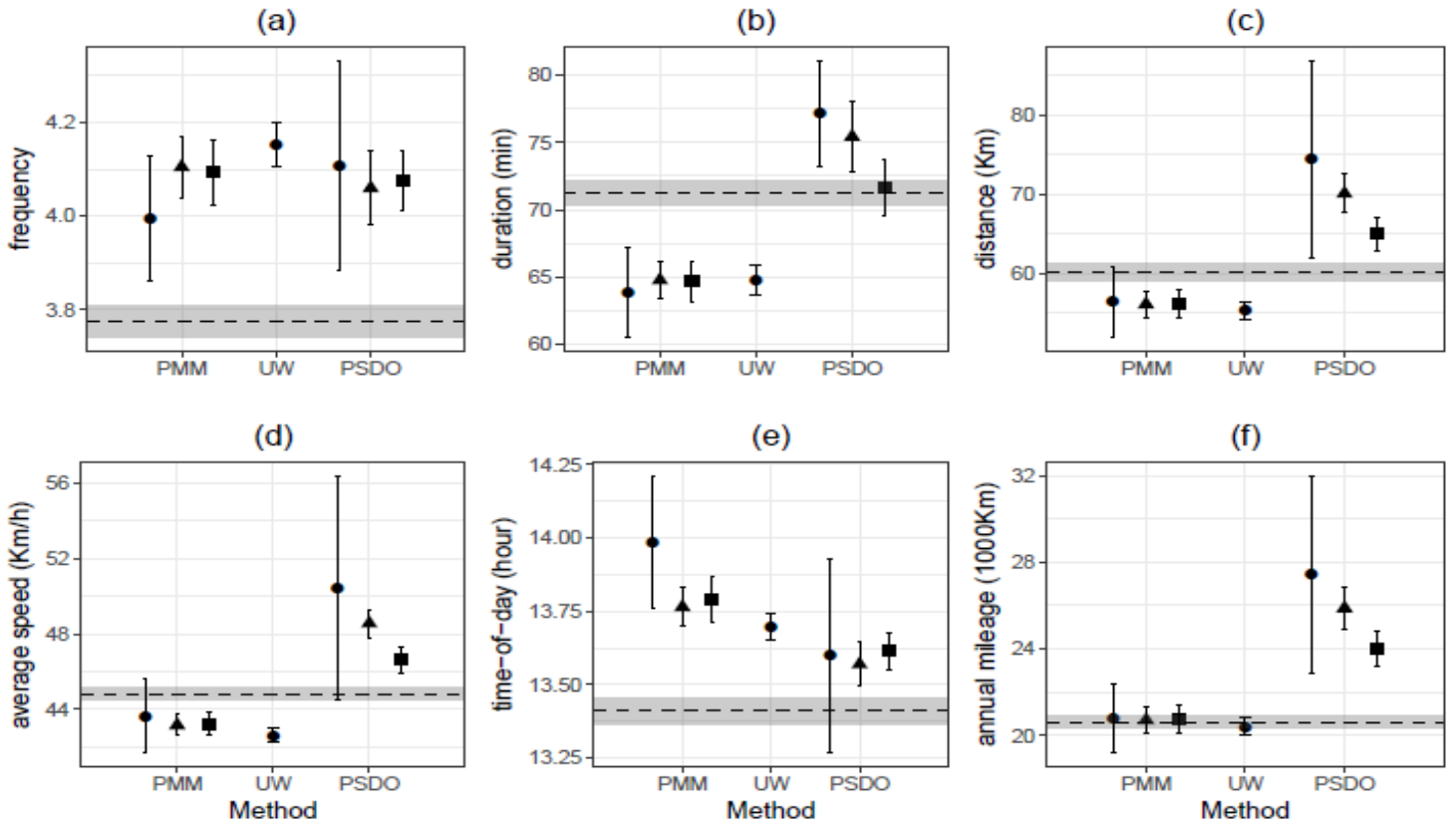


Kernel density of continuous covariates using pseudo-weights:



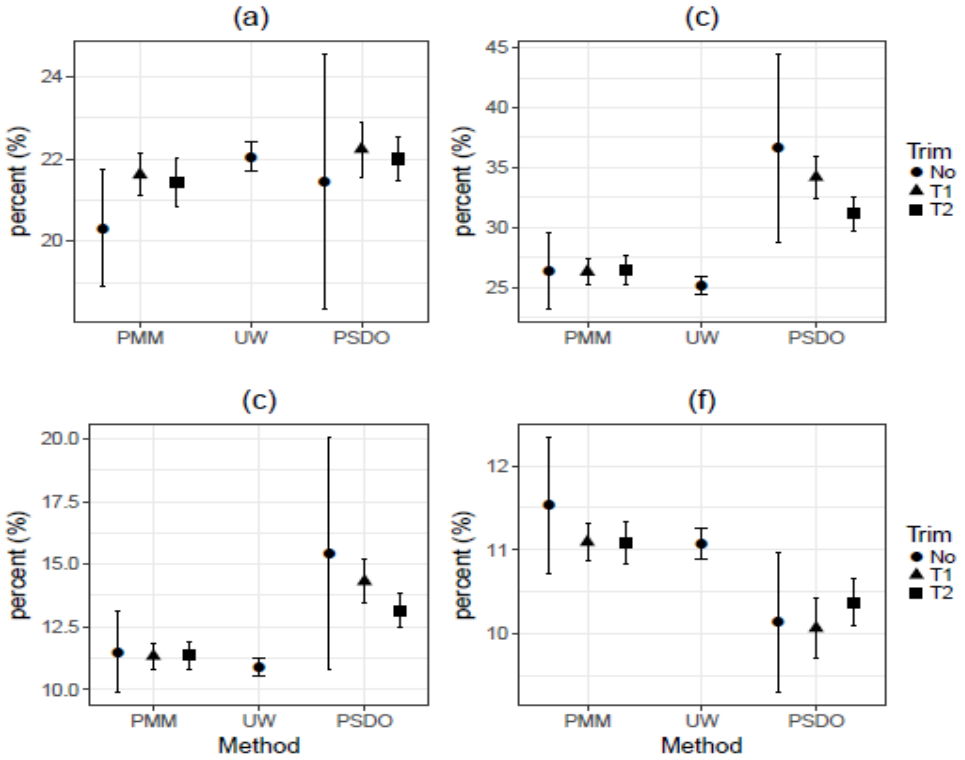
Weighted estimates of some trip-related outcomes

We compare weighted estimates of some trip-related outcomes with NHTS using pseudo-weights and PMM: Mean daily (a) frequency, (b) total duration, (c) total distance, (d) average speed, (e) start time; (f) annual mileage.



Weighted estimates of some SPMD-specific outcomes

We also estimated some SPMD-specific outcomes: Mean percent (a) trips using interstate, (b) of total trip spent on interstate, (c) time spent stopped during trip, (d) trip started between 6 and 10 am.



Simulation

- Two correlated sets of covariates were generated: W associated with probabilities of selection, and X associated with the outcome of interest.

$$\begin{pmatrix} W_1 \\ W_2 \\ X_1 \\ X_2 \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & -\rho/2 & \rho & -\rho/2 \\ -\rho/2 & 1 & -\rho/2 & \rho \\ \rho & -\rho/2 & 1 & -\rho/2 \\ -\rho/2 & \rho & -\rho/2 & 1 \end{pmatrix} \right)$$

- Let Y_c and Y_b are two continuous and binary outcome variables as below:

$$Y_c|X = x \sim N(-2 + x_1 - 2x_2 + 3x_1x_2, 1)$$

$$Y_b|X = x \sim b\left(\frac{e^{-2+x_1-2x_2+3x_1x_2}}{1 + e^{-2+x_1-2x_2+3x_1x_2}}\right)$$

- Each units in the population were assigned two sets of unequal probabilities of selection, which were correlated with W through a *logistic* link as below:

$$P(S_i = 1|W) = \frac{e^{-2-2w_1+w_2+0.5w_1w_2}}{2(1 + e^{-2-2w_1+w_2+0.5w_1w_2})}$$

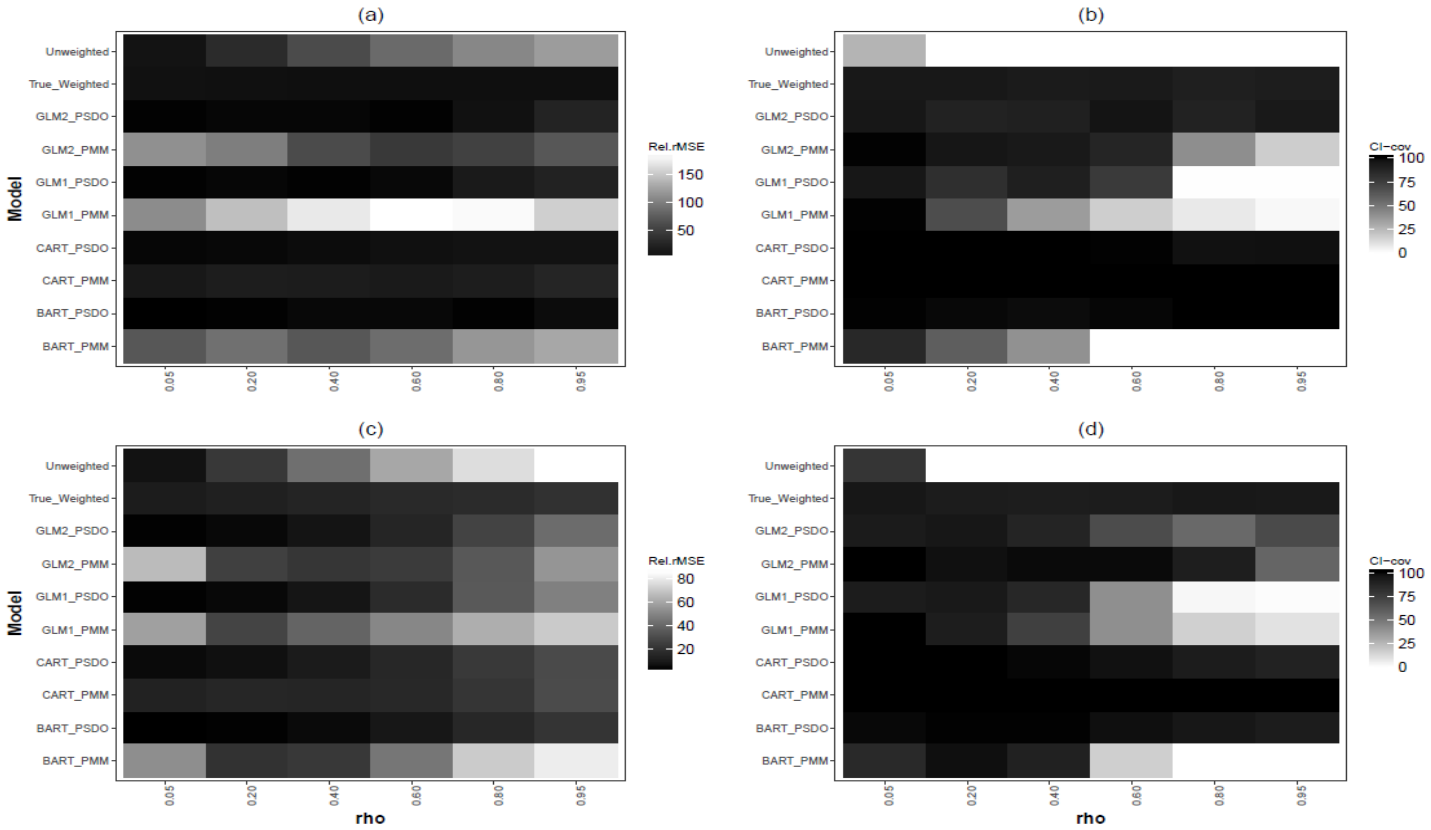
$$P(S_i^* = 1|W) = \frac{e^{-1+0.5w_1^2-w_2}}{4(1 + e^{-1+0.5w_1^2-w_2})}$$

Simulation

- The simulation was iterated 500 times, and relRMSE (standardized by mean) and nominal coverage rate of 95%CI were computed.
- Four models were utilized: (1) GLM1 (without interactions), (2) (1) GLM2 (with interactions), (3) BART, (4) CART, crossed with (a) pseudo-weight and (b) PMM construction.
- ρ=(0.05, 0.25, 0.50, 0.75, 0.95).

Simulation

- (a) and (c): relMSE for continuous and binary; (b) and (d) coverage rate for continuous and binary.



Discussion

Conclusion:

- Model specification seems to play a key role in the calibration process.
- BART works considerably better than alternative models in terms of bias reduction.
- Pseudo-weighting outperforms PMM, especially when bias is large.
- However, it may result in outliers, which need to be detected and trimmed properly.

Suggestions:

- Particular attention should be paid to identify and collect adequate predictive covariates.
- Pseudo-weighting with BART is recommended to be used for selection bias adjustment.

Next Steps

- Model based weight trimming might be helpful because of the unstable weights (Elliott and Little 2000; Elliott 2007, 2008)
- Data harmonization and measurement error: potentially differing definitions of "trips" in NHTS and SPMD
- Super-population modeling: generating synthetic populations from survey data for combination with non-probability data
- Sensitivity to failures of ignorability.

References

Elliott, M., Valliant, R. (2017)

Inference for nonprobability samples

Statistical Science 32(2), 249–264.

Chipman, H., George, E., McCulloch, R. (2007)

BART: Bayesian additive regression trees

The Annals of Applied Statistics 4(1), 266–298.

Kreuter, F., Peng, R. (2014)

12 Extracting Information from Big Data: Issues of Measurement, Inference and Linkage

Privacy, Big Data, and the Public Good: Frameworks for Engagement 257.

Elliott, M., Resler, A., Flannagan, C., Rupp, J. (2010)

Appropriate analysis of CIREN data: Using NASS-CDS to reduce bias in estimation of injury risk factors in passenger vehicle crashes

Accident analysis and prevention 42(2), 530–539.

Elliott, M. (2009)

Combining data from probability and non-probability samples using pseudo-weights

Survey practice 2(6).

References

Elliott, M., Little, R. (2000)

Model-based alternatives to trimming survey weights

Journal of Official Statistics 16, 191–209.

Elliott, M. (2007)

Bayesian weight trimming for generalized linear regression models

Survey Methodology 33, 23–34.

Elliott, M. (2008)

Model averaging methods for weight trimming

Journal of Official Statistics 24, 527–540.

References

Tan, Y.V., Flannagan, C., Elliott, M. (2017)

Predicting human-driving behavior to help driverless vehicles drive: random intercept Bayesian Additive Regression Trees

arXiv preprint arXiv 1609.07464.

Santos, A., McGuckin, N., Nakamoto, N.Y., Gray, D., Liss S. (2011)

Summary of travel trends: 2009 national household travel survey

Narla, S. (2013)

The evolution of connected vehicle technology: From smart drivers to smart cars to self-driving cars

ITE Journal 42(2), 530–539.

Kapelner, A., Bleich, J. (2013)

bartmachine: Machine learning with bayesian additive regression trees

arXiv preprint arXiv 21312.2171.

Questions?

Email: mrelliot@umich.edu