Assessing the use of multiple sources of auxiliary data for tailored survey designs

Ai Rene Ong, Rebecca Medway, Carol Tzu-Jou Wan

(American Institutes for Research)

ITSEW, Sept 18-20, 2024

Background and motivation

- Surveys using address-based samples (ABS) often do not have much information about the sampled households before the survey starts
- Auxiliary data can be used to provide additional information to tailor survey operations (e.g., targeting hard-to-reach populations), and to adjust survey estimates at the end of the data collection
- However, the usefulness of auxiliary data in survey sampling depends on the auxiliary data having sufficient quality and coverage of the frame
- We assessed two commercial data sources appended to National Household Education Survey (NHES:2023) for their quality and coverage

NHES:2023

- Cross-sectional nationally representative survey of households about early childhood care and education sponsored by the National Center for Education Statistics
- Sequential mixed-mode survey design with a web-push data collection protocol
- Two topical questionnaires: Early Childhood Program Participation (ECPP), Parent and Family Involvement in Education (PFI)
- NHES: 2023 has data from one commercial auxiliary data provider appended previously *(Vendor 1, Aux Data 1),* and also appends publicly available Census data.
- In this assessment of the quality of the auxiliary data, addresses from the NHES: 2023 frame were sent to another commercial vendor to match with their data *(Vendor 2, Aux Data 2)*
  - Note: Selected Aux Data 1 variables were used in sampling

Assessments of quality

- We examined the match rate and quality of the appended data for 205,000 of the addresses which were *sampled* for NHES:2023
  - About 86% of the addresses matched on both auxiliary data
- Following that, we examined the utility of Aux Data 2 in improving prediction accuracy for two key outcomes:
  - Response propensity (RP)
  - Presence of NHES-eligible children in the household (PC)

Sample sizes for the RP and PC models

- RP models:
  - NHES:2023 had various experiments with contact material, incentive structure, and mode offered.
  - Restricted to cases that can be matched to both Aux Data 1 and Aux Data 2 *and* baseline condition only (about 45,180 cases)
- PC models:
  - Restricted to respondents only regardless of experimental condition (about 98,650 cases)

Candidate variables (1)

- **Census Planning Database (PDB) variables (39 variables):**
  - Publicly available tract and block-group level American Community Survey (ACS) estimates and Census estimates
- **Aux data 1 variables (25 variables):**
  - Address-level demographic variables (e.g., age, gender), address-type variables (e.g., route type), and a few geographic-level variables (e.g., race/ethnicity)

All candidate predictors were recoded so that cases which were missing data for a given predictor were placed into an explicit "missing" category.

Candidate variables (2)

- **Aux data 2 variables (99 variables):**
  - 190 Variables available at different levels *(individual, household, address, geographic)*
  - Contains reported data *(e.g., number of children in the living unit)*, and modelled data *(e.g., likely to have grandparents in the household)*
  - Similar variables were collapsed into a single variable
  - Variables can be categorized into five types:
    - Presence of children/parents *(e.g., any children aged 0-18 in the living unit, mother or father is present in the household)*
    - Parenting/childcare products purchases *(e.g., likely to purchase toys at a retail location, frequent purchaser of baby products)*
    - Related to response propensity *(e.g., renter status, mail objectors)*
    - Mail and internet access *(e.g., likely to contain a parent who is frequently online, online shopper)*
    - Other occupant characteristics *(e.g., likely to include an AARP member, has an active military member)*

All candidate predictors were recoded so that cases which were missing data for a given predictor were placed into an explicit “missing” category.

Model building (1)

Model 1: PDB + Aux Data 1 (Baseline model)

- This model started with all the candidate variables from PDB and Aux Data 1, as was used in the variable selection process for the NHES:2023 RP and PC models.

Model 2: PDB + Aux Data 1 + Aux Data 2

- This model added the candidate variables from the Aux Data 2, in addition to the variables in Model 1.

Model building (2)

1. Conditional random forest (*cforest* function from the *partykit* package in R):
   - Used to narrow down candidate predictors for RP and PC models
   - 500 trees
   - Split criteria:
     - p-value: 0.05
     - Min. observations in terminal nodes: 120
2. Logistic regression with the variables selected from (1)
   - F-test testing full model with all selected variables from (1) vs. model with one variable removed
   - Only main effects were considered

Model fit statistics (3)

- In-sample and out-of-sample fit (k-fold):
  - Correct classification rate
  - True positive rate
  - True negative rate
  - Area under the curve (AUC)
- In-sample fit only:
  - Mc Fadden’s pseudo R-squared

RP Model: Predictors selected

Variable	Category	Coefficient	Std. Err.
The household uses internet service	1: Yes	0.17**	0.030
Likely to be in the market for a Student Loan in the next 180 days	0: No	-0.17*	0.073
	1: Yes	-0.32**	0.074
Buy by mail in at least one category (clothes, gardening, gifts, books, children's products)	1: Yes	0.07*	0.029
Donor to private foundations	1: Yes	0.11**	0.029
Match level of the Aux Data 2	H: Household	0.05	0.031
	P: Person	0.15**	0.028
Frequent mail order buyer of books or family/general magazines	0: No	0.10**	0.030
	1: Yes	0.18**	0.041

**p < 0.01, *p < 0.05, ^p < 0.10; reference category for all variables is “Missing”.

SOURCE: U.S. Department of Education, National Center for Education Statistics, NHES, 2023. Aux data 1 and 2.

RP Model: In-sample fit statistics

	Model 1	Model 2
Correct classification rate	64.6%	64.8%
True positive rate	62.1%	61.8%
True negative rate	67.7%	68.4%
Area under the curve (AUC)	0.703	0.707
McFadden’s pseudo R-squared	0.095	0.098

SOURCE: U.S. Department of Education, National Center for Education Statistics, NHES, 2023. Aux Data 1 and 2.

RP Model: Out-of-sample fit statistics

	Model 1	Model 2
Correct classification rate	64.4%	64.6%
True positive rate	65.4%	64.3%
True negative rate	63.2%	64.9%
Area under the curve (AUC)	0.695	0.699

SOURCE: U.S. Department of Education, National Center for Education Statistics, NHES, 2023. Aux Data 1 and 2.

PC Model: Predictors selected (1)

Variable	Category	Coefficient	Std. Err.
Likely to have an elderly person	1: Yes	-0.14 **	0.024
Likely a high spender or frequent purchaser at children's stores during the holiday season	0: No	0.05	0.054
	1: Yes	0.18 **	0.057
Viewers of family films	0: No	-0.14	0.718
	1: Yes	0.01	0.718
Purchased by direct mail through multiple companies	1: Yes	-0.20 **	0.028
Buyer of children or parenting products	1: Yes	0.50 **	0.025
Presence of child aged 0-18	0: No	0.71	0.627
	1: Yes	0.81	0.627

SOURCE: U.S. Department of Education, National Center for Education Statistics, NHES, 2023. Aux Data 1 and 2.

PC Model: Predictors selected (2)

Variable	Category	Coefficient	Std. Err.
Likely to compare prices across different sites before purchasing and typically read online reviews and consumer reports	1: Yes	0.07 **	0.019
Likely to include AARP members	0: No	0.28	0.623
	1: Yes	0.15	0.624
Likely to include child ages 0-18	1: Yes	0.18 **	0.029
Donor to education charities	1: Yes	0.10 **	0.024
Frequent mail order buyer of books or family/general magazines (buys at least 3 times through mail order)	0: No	-0.06 ^	0.032
	1: Yes	-0.15 **	0.041
High spender on children's products (online or in-store)	1: Yes	-0.12 **	0.025
Purchased children's apparel or merchandise in 2018-2023	1: Yes	0.46 **	0.046

SOURCE: U.S. Department of Education, National Center for Education Statistics, NHES, 2023. Aux Data 1 and 2.

PC Model: Predictors selected (2)

Variable	Category	Coefficient	Std. Err.
Likely to spend a quiet evening at home rather than go out; time is more valuable than money; duty before enjoyment and spending time at home with family	1: Yes	0.10 **	0.020
Spanish language preference for the person at the address, based on Aux Data 2's predictive name analysis	0: Non-Spanish	0.05	0.036
	1: Spanish	0.17 **	0.054
Individual's political affiliation	Democrat	-0.20	0.373
	Independent	-0.17	0.373
	Non-registered	-0.07	0.374
	Republican	-0.12	0.373
Likely to be brave/courageous	1: Yes	0.08 **	0.023
Likely to be affectionate/ passionate	1: Yes	-0.08 **	0.020
Buy by mail in at least one category (clothes, gardening, gifts, books, children's products) or prefers to shop by mail	1: Yes	-0.22 **	0.024

SOURCE: U.S. Department of Education, National Center for Education Statistics, NHES, 2023. Aux Data 1 and 2.

PC Model: In-sample fit statistics

	Model 1	Model 2
Correct classification rate	74.4%	75.8%
True positive rate	80.1%	78.1%
True negative rate	71.9%	74.8%
Area under the curve (AUC)	0.834	0.841
McFadden’s pseudo R-squared	0.263	0.277

SOURCE: U.S. Department of Education, National Center for Education Statistics, NHES, 2023. Aux Data 1 and 2.

PC Model: Out-of-sample fit statistics

	Model 1	Model 2
Correct classification rate	74.1%	75.2%
True positive rate	79.4%	78.6%
True negative rate	71.8%	73.8%
Area under the curve (AUC)	0.829	0.837

SOURCE: U.S. Department of Education, National Center for Education Statistics, NHES, 2023. Aux Data 1 and 2.

Summary and future work

- Only a few variables were selected from Aux Data 2 for the RP model 2, which suggests that Aux Data 2 may not add beyond what is already possible with PDB and Aux Data 1.
  - Fit statistics were similar across the model without Aux Data 2 and with Aux Data 2.
- More variables were selected from Aux Data 2 for the PC model 2, possibly due to many more variables available which are related to the presence of children in the household
  - Model with Aux Data 2 was able to identify households without children a little more accurately than the model without Aux Data 2 but did similarly to the model without Aux Data 2 in identifying households with children
- Future work: Other variable selection methods, other outcomes of interest (e.g., bilingual households)

AIR®

Advancing Evidence.
Improving Lives.

Thank you

aong@air.org

Appendix

Advancing Evidence.
Improving Lives.

RP: Selected variables for Model 1 (PDB)

Variable description	Frequency of selection
Self-response rate in the tract in the American Community Survey (ACS)	0.910
The predicted likelihood that the tract will produce a low mail return rate	0.616
Percentage of completed 2010 Census mail forms received from addresses in a mailback type of enumeration area	0.468
Percentage of ACS population in the tract that indicates no Hispanic origin and only race as “Black, African American, or Negro” or reports entries such as African American , Kenyan, Nigerian, or Haitian	0.460
Percentage of ACS population in the tract aged 25 years or over that has a college degree or higher	0.394

SOURCE: U.S. Department of Education, National Center for Education Statistics, NHES, 2023. Aux Data 1 and 2.

RP: Selected variables for Model 1 (Aux Data 1)

Variable description	Frequency of selection
Age of head of household	0.934
Person 2 age	0.916
Owner/renter status of head of household	0.896
Anyone 65 or older	0.826
Collapsed ethnicity of head of household	0.824
Phone type available on frame	0.686
Anyone 35-64	0.472
Bilingual mailing flag	0.450
Adult count in household	0.334
Education of head of household	0.314
Collapsed household income	0.306

SOURCE: U.S. Department of Education, National Center for Education Statistics, NHES, 2023. Aux Data 1 and 2.

RP: Selected variables for Model 1 (Aux Data 2)

Variable description	Frequency of selection
Uses an internet service	0.498
Likely to be in the market for a Student Loan in the next 180 days	0.446
Buy by mail in at least one category (clothes, gardening, gifts, books, children's products) or prefers to shop by mail	0.396
Donor to private foundations	0.376
Data match level	0.368
Frequent mail order buyer of books or family/general magazines (buys at least 3 times through mail order)	0.334

SOURCE: U.S. Department of Education, National Center for Education Statistics, NHES, 2023. Aux Data 1 and 2.

PC: Selected variables for Model 1 (PDB)

Variable description	Frequency of selection
Percentage of all ACS occupied housing units in the tract where one or more people are ages 18 years or under	0.878
Average number of persons per ACS occupied housing unit in the tract. Calculated by dividing the total household population from the ACS in the tract by the total number of occupied housing units from the ACS in the tract	0.794
Percentage of all ACS occupied housing units in the tract where the householder and his or her spouse are listed as members of the same household; does include same- sex married couples	0.692
Median ACS household income for the block group	0.384
Percentage of ACS population in the tract aged 25 years or over that has a college degree or higher	0.374
Self-response rate in the tract in the American Community Survey (ACS)	0.356
Percentage of ACS occupied housing units in the tract that are not owner occupied, whether they are rented or occupied	0.328

SOURCE: U.S. Department of Education, National Center for Education Statistics, NHES, 2023. Aux Data 1 and 2.

PC: Selected variables for Model 1 (Aux Data 1)

Variable description	Frequency of selection
Adult count in household	1.000
Age of head of household	1.000
Person 2 age	1.000
Anyone 65 or older	0.986
Anyone 35-64	0.962
Child count in household	0.930
Marital status of head of household	0.900
Collapsed ethnicity of head of household	0.848

SOURCE: U.S. Department of Education, National Center for Education Statistics, NHES, 2023. Aux Data 1 and 2.

PC: Selected variables for Model 1 (Aux Data 2)

Variable description	Frequency of selection
Likely to have an elderly person	0.902
Likely a high spender or frequent purchaser at children's stores during the holiday season	0.892
Viewer of family films	0.868
Purchased by direct mail through multiple companies	0.856
Children/parenting products	0.846
Likely to include age child ages 0-181	0.844
Like to compare prices and read online reviews across different sites before purchasing	0.774
Likely to include an AARP member	0.758
Likely to include age child ages 0-181	0.752
Donor to education charities	0.674
Frequent mail order buyer of books or family/general magazines	0.634

SOURCE: U.S. Department of Education, National Center for Education Statistics, NHES, 2023. Aux Data 1 and 2.

PC: Selected variables for Model 1 (Aux Data 2) – Cont’d

Variable description	Frequency of selection
Household is a high spender on children's stores	0.594
Likely to purchase products categorized as "Children's Merchandise" or "Children's Apparel" again	0.572
The most recent purchase in the household, in the category "Children Merchandise" or "Children Apparel."	0.548
Likely to spend a quiet evening at home than go out; time is more valuable than money, duty before enjoyment and spending time at home with family.	0.484
Spanish language preference for the person at the address, based on predictive name analysis	0.436
Individual's political affiliation	0.434
Likely to be brave, courageous, daring, adventuresome, broadminded, open-minded, liberal, tolerant, creative, inventive, imaginative, artistic, funny, humorous, amusing, witty, intelligent, smart, bright, and well informed.	0.422
Likely to be affectionate, passionate, loving, romantic, amicable, amiable, affable, benevolent, kind, good-hearted, warmhearted, sincere, sociable, friendly, cheerful, likable, trustworthy, competent, reliable, and responsible.	0.418
Buy by mail in at least one category (clothes, gardening, gifts, books, children's products) or prefers to shop by mail	0.326
Annual predicted discretionary spend for education	0.326

SOURCE: U.S. Department of Education, National Center for Education Statistics, NHES, 2023. Aux Data 1 and 2.