THE NATIONAL CHILDREN'S STUDY

HEALTH GROWTH ENVIRONMENT

A Comparative Evaluation of Traditional Listing vs. Address-Based Sampling Frames

Valerie Hsu
Jill M. Montaquila
J. Michael Brick

Westat

March 17, 2010

Overview

1. The National Children's Study
2. Research goals
3. Evaluation
   - Overview of evaluation design
   - Data quality control checks and matching procedures
   - Results of matching
   - Modeling match rates
4. Discussion
   - Summary of findings
   - Future research

The National Children's Study

- Designed to examine the effects of environmental influences on the health and development of about 100,000 children across the U.S., following them from before birth until age 21
- Multi-stage area probability household sample
  - Primary sampling units (PSUs): Typically single counties (about 10,000 addresses per PSU)
  - Segments: Clusters of contiguous census blocks (typically about 500-1200 households per segment)
  - All births in sampled segments are eligible; household-based data collection

The National Children's Study (Cont.)

- This evaluation is based on listing conducted for a Pilot Study in seven PSUs:

PSU	Location
DC	Duplin County, NC
BYPL	Brookings County, SD; Yellow Medicine, Pipestone, and Lincoln Counties, MN
WC	Waukesha County, WI
MC	Montgomery County, PA
SLC	Salt Lake County, UT
OC	Orange County, CA
QC	Queens County, NY

Our Research

- Comparisons between traditionally listed addresses and geocoded USPS-based addresses
- Gaining an understanding of whether particular kinds of places are more likely to be undercovered
- An approach for assessing when USPS lists could be used in place of traditional listing based on a "match rate" model

Overview of Evaluation Design

- **Traditional listing:**
  - **Listers used hard-copy forms to record addresses within sampled segments**
  - **Different listers in each PSU**
- **USPS-based address lists geocoded to blocks in sampled segments**
- **Matching of the two lists**

Data Quality Control Checks and Matching Procedures

- Automated exact matching
- Manual matching to resolve:
  - Differences in spelling/typos (e.g., "Weatherby Rd." vs. "Wetherby Rd.")
  - Differences in street type (e.g., "Oak St." vs. "Oak Ln.")
  - Other variations in street specifications (e.g., "23rd St." vs. "23 St.")
  - "No number" addresses (e.g., matching a "no number" address listed between 123 Main St. and 127 Main St. with a "125 Main St." listing on the USPS-based list)

Data Quality Control Checks and Matching Procedures (Cont.)

- A few blocks/apartment complexes that were missed completely by the listers were identified during matching process
- Listers were sent out to relist the block(s) in question:
  - Two segments (92 additional addresses) in BYPL;
  - One segment (12 additional addresses) in WC;
  - One segment (42 additional addresses) in OC; and
  - Two segments (70 additional addresses) in MC.
- Augmented traditional listing contained traditionally listed addresses with corrections less addresses listed in error

Match Rate

- Assume that the augmented traditional listing list is the "gold standard"
- Match rate calculation:

$$\frac{\#\text{ addresses on both lists}}{\#\text{ addresses on augmented traditional listing list}}$$

- The proportion of traditionally listed addresses that would have been obtained from a USPS list

Results (Cont.)

PSU-Level Matching Results

PSU	Urbanicity (%)	Match rate (%)	Matches obtained through manual matching (%)	Unmatched USPS addresses (%)
DC	14	50	17	23
BYPL	44	54	25	13
WC	88	91	11	5
MC	97	86	13	6
SLC	99	92	6	3
OC	100	96	6	1
QC	100	94	34	2

Results (Cont.)

Nongeocodables and Multi-Drops

PSU	Urbanicity (%)	Nongeocodable USPS-based addresses (%)	Multi-drop USPS addresses (%)
DC	14	18	0.10
BYPL	44	25	0.05
WC	88	5	0.14
MC	97	4	0.26
SLC	99	7	0.03
OC	100	2	0.03
QC	100	<1	10.42

Results (Cont.)

- Example of matching multi-drop addresses:
  - USPS List

Street No.	Street Name	Street Type	Unit No.	Unit Type	Drop Count
123	Main	St			3

  - Traditionally Listing List

Street No.	Street Name	Street Type	Unit No.	Unit Type
123	Main	St	Apt	A
123	Main	St	Apt	B
123	Main	St	Apt	C

Results (Cont.)

- There is variation in match rates at the segment level (e.g., match rates ranging from 21% to 92% in BYPL, and from 72% to 100% in OC)
- Beneficial to identify (a priori) areas where USPS lists could be used in lieu of traditional listing

Modeling Match Rates

- Predicting match rates (i.e., coverage rates of the USPS lists relative to what might be expected from traditional listing) using multiple regression
  - Explored relationships between segment characteristics and match rates
  - Used selected statistics from the Census 2000 Summary File 1, Summary File 3, and 2005-2007 ACS to build a prediction model
    - Variables such as proxy for new housing development, measures of stability of occupancy, and classification of types of structures

Final Model

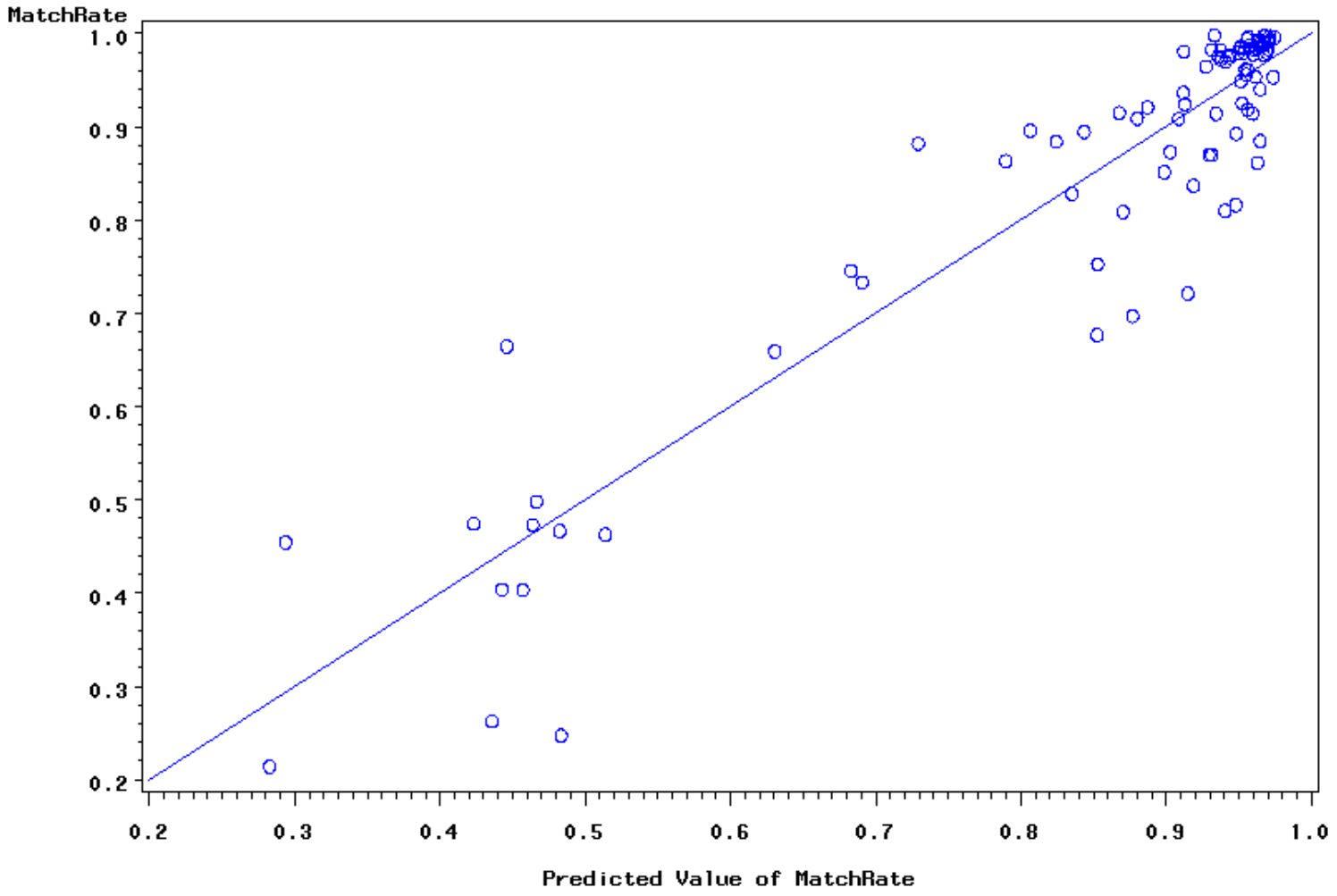
- Final model used to predict match rates:

Predictor	Parameter Estimate	Standard Error
INTERCEPT	-0.79**	0.27
SAME_HU_LSTYR	-0.01	0.22
BUILT_05LTR	28.04***	4.53
URBAN	0.77***	0.07
OCCUPIED	1.02***	0.26
BUILT_05LTR* URBAN	-28.91***	5.09

****p-value*<0.001 ***p-value*<0.01 **p-value*<0.05

- Final model fit the data adequately (F=107.71 and R^2=0.86)
- Urbanicty had the greatest effect on match rates

Match Rates vs. Model Predicted Rates



Using Match Rate Models to Decide Which Areas to List

- Determine an operationally efficient match rate threshold that defines adequate coverage
  - If a segment has a predicted match rate that falls below the specified threshold, traditional listing is used; otherwise, USPS lists are used
- We considered two threshold values, 0.7 and 0.8

Using Match Rate Models to Decide Which Areas to List (Cont.)

Model Performance Using Threshold Value = 0.7

		Predicted match rate		Total
		Below threshold	Above threshold	
Actual match rate	Below threshold	13 (87%)	2 (13%)	15
	Above threshold	2 (3%)	74 (97%)	76
	Total	15	76	91

Model Performance Using Threshold Value = 0.8

		Predicted match rate		Total
		Below threshold	Above threshold	
Actual match rate	Below threshold	15 (79%)	4 (21%)	19
	Above threshold	2 (3%)	70 (97%)	72
	Total	17	74	91

Discussion

- Match rates at the PSU-level are:
  - Generally higher in urban areas than rural, and
  - Generally lower in areas with higher rates of new construction,

  BUT there was variation in match rates at the segment level
- Address lists may be used to check the traditional listings; extent of manual matching is a consideration
- Use of missed unit procedures to increase coverage of USPS lists

Discussion (Cont.)

- Important to consider the limitations of USPS lists and the consequences of using them as sampling frames
- Greater coverage of the USPS lists might be achieved in designs in which the sample is selected from a list frame

Discussion (Cont.)

- With respect to match rate model, cross-validation is necessary to ensure that over-fitting is not an issue
- As Census 2010 and additional ACS data become available, refitting the prediction model is useful
- Threshold should be set based on a variety of considerations (e.g., the skill and training of the listers, the effectiveness and cost of missed unit procedures, and the relative costs of traditional listing and USPS listings)

Future Research

1. Validation of "match rate" model using an independent set of segments
   - Re-fitting the model as new covariate data become available
2. Examination of eligibility of households/persons associated with the following types of addresses:
   - addresses on both the USPS and traditional listing lists,
   - addresses only on traditional listing lists, and
   - additional missed units added during data collection

Contact Information

Valerie Hsu

Westat

1600 Research Blvd.

Rockville, MD 20850

valeriehsu@westat.com