EVALUATING THE USE OF RESIDENTIAL MAILING ADDRESSES IN A METROPOLITAN HOUSEHOLD SURVEY

Vincent G. Iannacchione, Jennifer M. Staab, and David T. Redden
Research Triangle Institute

Abstract

We developed a sampling frame for a probability-based household survey by purchasing an exhaustive list of over 818,000 residential mailing addresses in Dallas County, Texas. The addresses were obtained from the Delivery Sequence File (DSF) offered by the US Postal Service (USPS) through a nonexclusive license agreement with private companies. The DSF is a computerized file that contains all delivery point addresses serviced by the USPS, with the exception of general delivery.

We used the geographic coordinates of the addresses to construct digital maps of the immediate vicinity around each selected address to help the field interviewers locate the selected address. To evaluate the coverage of the mailing addresses, we selected a sub-sample of 2,498 addresses and used the Half-Open Interval (HOI) procedure (Kish 1965) to search for missed housing units in the interval between the selected address and the next address in delivery sequence order. A total of 46 missed addresses were found with the HOI procedure. Also, we found that the vast majority of persons who maintain a residential P.O. Box also have mail delivered to their street address. The 90 percent occupancy rate is consistent with other metropolitan household surveys that use traditional on-site enumeration methods.

Key Words: Sampling Frame, Delivery Sequence File, Half-Open Interval Procedure,
Geographic Information Systems

Introduction

On-site enumeration requires field staff to visit selected geographic areas and then create an exhaustive listing of all potential housing units in each selected area. Although generally regarded as the most complete enumeration process available for household surveys, the time and expense associated with on-site enumeration precludes it from being a viable option for many household surveys. This paper describes our experience using residential mailing address lists instead of on-site enumeration procedures. We describe the advantages and disadvantages of using residential mailing address lists, and our use of the half-open interval linking procedure to estimate the residential mailing list's coverage of our desired population.

Design and Methodology

We used residential mailing lists to develop sampling frames for two metropolitan household surveys in Texas. The first study, conducted from January 2000 to July 2000 in Houston Texas, used a residential mailing list to enumerate the sampling frame and isolate the desired study population. The objective of the study was to estimate the impact of a mass media campaign upon the poorest adult African-American adults (age 18-45) living within the city limits. Due to the nature and timing of the media campaign, on-site enumeration of dwelling units was not a viable option. Instead, all mailing addresses within the poorest census tracts with highest density of

African-American population were used to define the sampling frame. From that frame, we randomly selected 2,724 houses to screen. Of those addresses, 2,635 (97%) were locatable/ non-demolished housing units. This high percentage was important in verifying that the mailing list did not contain a high percentage of non-existing housing. However, one major limitation of this implementation of the mailing address list to enumerate houses is that it did not contain any linking procedure to identify housing units not included on the residential mailing list. Therefore, coverage could not be evaluated.

The second household survey is currently being conducted in Dallas County and is the primary focus of this paper. The target population for this study of heart disease prevention consists of the estimated 1.25 million civilian, non-institutionalized adults, aged 18 to 65, who reside (or will reside) in Dallas County, Texas between July, 2000 and September, 2001 and who speak English or Spanish. To enable comparisons between ethnic subgroups, minority communities in the County were over-sampled. During the 15 months of data collection, we expect to select 15,000 addresses for screening and to complete interviews with approximately 6,100 eligible persons. We selected a sub-sample of 2,498 addresses and used a linking procedure to identify residential addresses not included on the mailing list and used the results to evaluate how well the mailing lists cover the residential population of Dallas County. The remainder of this paper describes how we constructed the sampling frame and how we designed and implemented the linking procedure to identify addresses not included on the frame.

Construction of the Sampling Frame

We began construction of the sampling frame by purchasing the entire list of residential mailing addresses in Dallas County (over 818,000 addresses excluding residential P.O. Boxes) based on the Delivery Sequence File (DSF), a service offered by the US Postal Service (USPS) through a nonexclusive license agreement with private companies. The DSF is a computerized file that contains all delivery point addresses serviced by the USPS, with the exception of general delivery. While the DSF is essentially a complete list of residential addresses, the private companies who offer the addresses must delete the addresses of persons who request to have their addresses taken off the list. These deletions represent a source of under-coverage on the sampling frame.

The low cost associated with purchasing mass-mailing lists enabled us to obtain the entire list of residential addresses in Dallas County. (In fact, the cost of the entire list was at least ten times less than the estimated cost of on-site enumeration for selected areas within the County.) Obtaining all mailing addresses enabled us to reduce design effects by dispersing the sample throughout the County instead of restricting it to a number of small geographic clusters to facilitate on-site enumeration. It also facilitated the creation of custom maps of neighborhoods surrounding each sample address. Finally, having the entire set of addresses enabled us to identify potential gaps in the delivery sequence of postal carrier routes.

The DSF contains the delivery sequence number of addresses on each postal carrier route. The delivery sequence number identifies the order in which the mail is delivered. We identified 11,644 addresses (1.4 percent) on the sampling frame that are associated with gaps in the delivery sequence number. As we describe later in the paper, these gaps appear to be correlated with the occurrence of adjacent addresses that are not on the mailing lists. We speculate that many of these missing addresses are for persons who requested to have their names removed from the mass mailers' lists.

After purchasing the mailing list, the mailing addresses were geocoded to determine geographic coordinates (i.e., latitude and longitude). The standardized format of the mailing addresses on the DSF enabled us to obtain geographic coordinates for more than 99.4 percent of the purchased addresses. We used the geographic coordinates to move from "postal geography" (i.e., zip codes) to "Census geography" (i.e., Census Tracts and Blocks) so that we could use Census data to stratify the sample by race/ethnicity. The geographic coordinates also facilitated the creation of digital maps for all selected addresses.

As **Table 1** shows, we excluded a small number of purchased addresses from the sampling frame. These included addresses that were unable to be geocoded, a small number of multi-drop addresses (i.e., multiple persons associated with the same address), and addresses in primarily commercial/industrial Census Tracts with fewer than 50 residential addresses. Finally, although we did not purchase them, residential P.O. boxes constituted the vast majority of addresses that were excluded from the frame. In spite of these exclusions however, the total number of residential addresses on the sampling frame compares favorably to the 1998 estimate of the number of occupied housing units in Dallas County.

Implementation of the Half Open Interval Linking Procedure

We implemented the Half-Open Interval (HOI) linking procedure (Leslie Kish 1965) on a randomly selected sub-sample of 2,498 addresses to estimate the amount of under-coverage associated with the sampling frame. The HOI procedure adds housing units to an existing frame by searching for new units in the interval between the selected unit and the next unit on the frame. The actual inspection of the frame needs to be done only within the selected intervals. New units that are discovered during field interviewing are automatically included in the sample.

To be effective, the HOI procedure requires that the addresses on the frame be sorted in geographically proximal order. We achieved this ordering by arranging the list in delivery sequence order, (i.e., the order in which the mail is delivered) within the 1,727 city and rural carrier routes identified on the sampling frame. As **Figure 1** shows, the delivery sequence on a postal carrier route usually proceeds up one side of a street and down the other making it very amenable to the HOI procedure. Based on our sample, we estimated that an HOI could be constructed and located for approximately 94 percent of the addresses on the frame.

The most common reasons why an HOI could not be constructed were:

- Address at the end of a contiguous portion of a carrier route; and
- Irregular delivery sequence (e.g., next address across the street).

We were able to locate all but 11 of the 2,380 HOIs constructed for the sub-sample.

In addition to addresses without an HOI, the effectiveness of the HOI procedure was adversely affected by the fact that 38 percent of the addresses on the sampling frame were apartments. Within an apartment complex, the HOI procedure reduces to checking for missed addresses between individual apartments. As a result, the HOI of the last address associated with an apartment complex is usually the only interval likely to include a missed address.

Table 1. Exclusions from the Sampling Frame of Residential Mailing Addresses

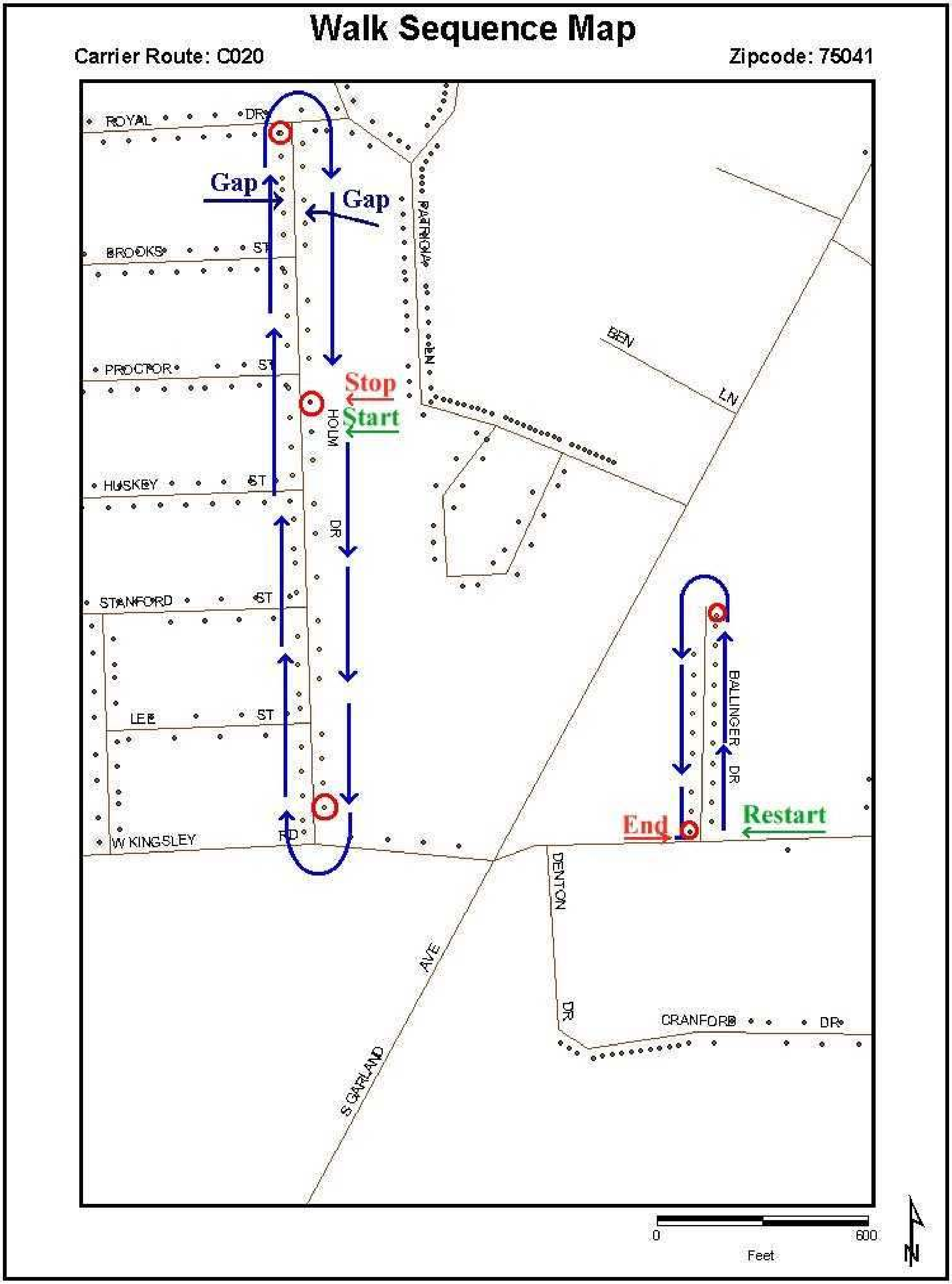
	City Routes	**Rural Routes**	**All Routes**
Number of addresses in Dallas County[1]	**768,612**	**73,331**	**841,943**
Exclusions:			
Residential P.O. Box	18,388	5,130	23,518
Address not able to be geocoded	4,775	627	5,402
Multi-drop address[2]	1,046	0	1,046
Census Tracts with fewer than 50 addresses	962	0	962
Number of addresses on sampling frame	**743,441** (96.7%)	**67,574** (92.2%)	**811,015** (96.3%)
Number of occupied housing units[3]	-	-	763,492

[1] Addresses for city postal routes were purchased from ADVO, Inc.; addresses for rural postal routes were purchased from Donnelley Marketing Services.

[2] Multiple persons associated with the same address.

[3] 1998 estimate from the North Central Texas Council of Governments.

Figure 1. Example of a Delivery Sequence for a Postal Carrier Route



Results

Prevalence of missed addresses: We identified 46 missed addresses among the 2,369 HOIs constructed and located for the sub-sample. All but three of the missed addresses appeared to be occupied. The missed addresses were dispersed throughout the County and appeared to follow the population distribution. Only two of the HOIs in the sub-sample yielded more than one missed address. Nine missed addresses were found among the 986 apartment HOIs. Even though the mailing lists used to construct the sampling frame were about a year old, only five of the missed addresses appeared to be new construction. The (one-sided) 95% confidence limit for the prevalence of missed addresses on the sampling frame is 2.5 percent.

Relationship between gaps in the delivery sequence and missed addresses: As **Table 2** shows, there were 51 addresses in the sub-sample that were associated with gaps in the delivery sequence. Of these, 34 (67%) were adjacent to a missed address. The agreement between the occurrence of a gap and a missed address (as measured by Cohen's kappa and McNemar's test) is at least moderately significant.

Table 2. Relationship between Gaps in the Delivery Sequence and the Occurrence of Missed Addresses

Did the sample address have:			
A gap in the delivery sequence?	**At least one missed address in the adjacent HOI?**		
	Yes	**No**	**Total**
Yes	34	17	51
No	8	2,310	2,318
Total	42	2,327	2,369
Cohen's kappa: 0.726 McNemar's Test: 3.24 P-Value: 0.0719			

Residential P.O. Boxes: We suspected that many, if not most, of the persons who maintain residential P.O. Boxes also receive mail at their street address. Therefore, we included a question on the household screening questionnaire that asked whether the sampled resident maintained a residential P.O. Box. Our current sample estimate of the total number of addresses associated with residential P.O. Boxes is 20,873 (s.e. of 2,699). This is very consistent with the county-level total of 23,518 residential P.O. Boxes and seems to confirm our suspicion that the vast majority of persons with residential P.O. Boxes also have mail delivered to their street address.

Conclusions and Recommendations

We believe these results demonstrate the utility and completeness of using mailing addresses to develop a sampling frame for a metropolitan household survey. While there are sources of under-coverage (e.g., addresses of persons who request that their names be removed from a mass-mailer's list), the use of a sampling frame linking procedure such as the HOI procedure can increase the coverage of the lists. In fact, the apparent relationship between the occurrence of gaps in the postal delivery sequence numbers can be used to increase the efficiency of the sampling design by identifying and possibly over-sampling addresses associated with gaps in the delivery sequence.

We caution that this evaluation is limited to one large metropolitan area. Further research is needed to examine the use of residential mailing addresses in other surveys, especially those that include rural areas.

Reference

Kish, Leslie (1965). *Survey Sampling*, John Wiley & Sons, New York. p 56.