USE OF GEOGRAPHICALLY CLASSIFIED TELEPHONE DIRECTORY LISTINGS IN MULTI-MODE SURVEYS

R. W. Whitmore, R. E. Mason, T. D. Hartwell, M. S. Rosenzweig
Research Triangle Institute

1. Introduction

The purpose of this paper is to discuss the use of telephone directory listings that are classified according to Census geographic variables in association with standard area household sampling techniques. These listings are currently available from such sources as Donnelley Information Services, Inc. and Survey Sampling, Inc. for the United States household population.

The Research Triangle Institute (RTI) recently used listings purchased from Donnelley Corporation in association with standard area sampling techniques. Census data tapes were utilized to select a first stage sample of area segments identified by Census block groups. A computer tape listing the selected block groups was then sent to Donnelley Corporation. The tape was returned with computerized listings of names, addresses, and telephone numbers for the selected block groups.

The listings purchased were compiled by Donnelley Corporation from two sources: (1) current telephone directory listings and (2) current vehicle registration records. Telephone directory listings are available for the entire United States. Vehicle registration records are available for about half the States. The telephone directory listings include name, address and telephone number information from current telephone directories. The telephone numbers in the listings purchased by RTI were unique, and listings that were obviously nonresidential, based on the name, had been removed. Some nonresidential listings remained, especially for doctors and lawyers. The listings based upon vehicle registration records contained only name and address information, not telephone numbers. To the extent possible, Donnelley Corporation added a single vehicle registration record for an address when there was no telephone directory listing for that address.

Conceptually, these lists can be used in place of the standard field lists of housing units to identify clusters of sample housing units. Moreover, the telephone directory listings can be used to conduct most of the interviews by telephone. If the target population contains households other than those with listed telephone numbers, as it most often does, field interviews are also required. However, field interviews are needed only for sample households not interviewed by telephone. This includes households that correspond directly to sample listings and so-called missed housing units. A missed housing unit is a housing unit that does not appear in the commercial listing frame and is linked to a sample housing unit via a unique linking rule, similar to a missed housing unit linking rule for a standard area household sample. Ideally, this use of the commercial listings can be expected to produce results comparable to standard area household sampling at reduced cost. Cost savings are realized since it is not necessary to send staff into the field to list all housing units in the selected block groups, and because most interviews can be conducted by telephone.

RTI used this type of procedure for two recent surveys. One was an EPA-sponsored study of personal exposure to carbon monoxide in the metropolitan areas of Washington, D.C., and Denver, Colorado. The other was a state-wide study of social service needs for the State of Louisiana. The sample for the CO study was primarily urban, while that for the Louisiana study was mostly rural. Between the two studies, there is a reasonable basis for assessing the usefulness of the commercial listings as sampling frames.

For both the CO study and the Louisiana study, the listings appeared to be reasonably complete. Most of the listings were also found to be correctly classified according to block group. Donnelley Corporation claims that their listings are about 95 percent complete. Our experience is not inconsistent with this claim. However, we found that the undercoverage does not seem to occur at random. Instead, both studies found that there were small geographic areas for which there were no listings whatsoever. The telephone directories for these areas simply had not been used in compiling the listings. Standard field procedures were used to list all housing units and select clusters of sample housing units for these block groups.

The major problem encountered in using the listings to identify sample housing units was that it was often difficult to locate the housing units corresponding to the sample listings in the field. The addresses generally came from telephone directory listings. Hence, most residents of apartment complexes all had the same address, namely the street address of the apartment complex. Similarly, households on rural roads tended to all have the same address, simply the name of the road on which they lived. This presented some problem for location of the sample housing units. But, more importantly, it made the check for missed housing units very difficult to implement correctly. Because of these problems, and other more subtle problems with the operational definition of missed housing units, we feel that there is no completely satisfactory way to perform the check for missed housing units for a sample from the commercial listings.

Based upon RTI's experience, recommendations with regard to use of geographically classified telephone directory listings in association with standard area household sampling techniques will be presented. Cost advantages of the proposed design relative to a standard area household design will be considered. Relevant RTI experience will then be discussed. In particular, actual telephone interview experience based upon sample listings will be presented. The quality of the commercial listings will also be addressed in more detail.

2. Design Recommendations

Based upon the cited experience using geographically-classified telephone directory

listings, the authors feel that such listings can best be utilized with a dual frame sampling procedure. In the dual frame approach, two independent samples are selected from the (complete) area frame:

(1) One sample is a standard area sample with sample clusters identified from field listings of all housing units in the selected area segments.
(2) The other sample uses the commercial listings to identify sample clusters in the selected area segments.

It is recommended that the commercial listing sample be used only to generate telephone interviews. Using this methodology, the standard area frame sample is used to compensate for the bias resulting from the telephone interviews generated by the commercial listing sample. In order to compensate for this bias, it is necessary to determine whether or not each household in the standard area frame sample is included on the commercial listing frame. This is easily done for commercial listings that come directly from the current telephone directory. A single questionnaire item can determine whether or not the household is served by a residential telephone number that is listed in the current telephone directory. It is not so easy to determine telephone coverage with respect to commercial listings based upon vehicle registrations. It would be necessary to determine whether or not each household contains an individual with a registered vehicle, and whether or not a current residential telephone number could be obtained for that individual from the information operator, based upon the name and address in the vehicle registration. Moreover, if vehicle registration records are included in the commercial listings frame only when there is no telephone number for an address, there may be a tendency to exclude vehicle registration records for residents of apartments and rural roads where the residents tend to all have the same address in the telephone directory listings. Thus, use of the vehicle registration listings makes bias correction using the standard area frame listings tenuous, at best. In conclusion, the authors recommend that only the telephone directory listings be used for the commercial listings sample. The standard area frame sample is used to compensate for the bias resulting from use of the frame of telephone directory listings.

Use of only the telephone directory listings for the commercial listing sample makes implementation of the dual frame methodology very straightforward. States with and without vehicle registration records in the commercial listings are handled in exactly the same way. For every sample household, one or two questionnaire items can be used to determine the number of residential telephone numbers that are listed in the current telephone directory for the household. This information is sufficient to facilitate unbiased estimation for linear statistics using either multiple frame multiplicity estimators, such as those discussed by Casady and Sirken [1980], or difference estimators, such as those discussed by Konijn [1973]. The difference estimators may be preferable since they address the bias correction more directly.

Both RTI projects using commercial listings found that there were some area segments with no commercial listings. In some cases, the listings may simply have been missclassified. In other cases, the listings actually were not available. Hence, a determination of whether or not telephone directory listings are available is needed for each area segment in the standard area household sample. If telephone directory listings are not available for some area segments in the standard area household sample, the households in these area segments must be treated for estimation as not represented in the frame of telephone directory listings. Otherwise, all households with a currently listed telephone number are treated as being present on the frame of telephone directory listings.

It is recommended that this dual frame approach be implemented as two half samples. This will generally be less costly than obtaining the same number of interviews from a standard area frame sample alone. Some savings will be achieved by using the commercial lists instead of lists of housing units produced by field staff to identfy sample clusters. The use of telephone interviews instead of field interviews may produce some additional savings.

This dual frame approach could, of course, also be used for a field half-sample and a random-digit-dial (RDD) telephone half-sample. Some advantages of using the geographically-classified telephone directory listings instead of random digit dialing for one half sample are the following:

(1) Census geographic variables can be used to oversample sub-populations of interest at the first stage.
(2) The proportion of telephone numbers called that are working residential numbers will be much higher for the telephone directory sample.
(3) If field follow-up interviews are necessary, such as for the EPA's personal monitoring studies, the geographic clustering will reduce subsequent field interview costs.

Of course, there is some loss in precision due to clustering and due to use of the incomplete telephone directory frame. These losses will generally be compensated by decreased cost for the sample survey. Thus, the proposed design is expected to be cost effective for certain types of studies.

3. EPA's CO Exposure Study

The sample design for the EPA's CO exposure study was a deeply stratified, three stage sample. The EPA purposively selected the metropolitan areas surrounding Washington, D.C., and Denver, Colorado, as the study sites. The purpose of the study was to monitor personal carbon monoxide (CO) exposure for residents of the study sites during the season with the highest CO levels. Area sample segments defined by Census geographic variables were selected at the first stage of sampling. Households were selected at the second stage, and all household members were administered a short screening interview. Persons were selected at the third

stage for personal CO monitoring. Only the first two stages of sampling are relevant to the topic of this paper, and the third stage sample of persons will not be discussed further.

The first stage sampling units (FSUs) were constructed using block groups and enumeration districts as defined for the 1980 Census to completely account for the geographic area of each study site. A sequential, minimum probability replacement (MPR) sampling procedure (See Chromy [1979]) was used to select the FSUs with probabilities proportional to the 1980 Census counts of occupied housing units. The frame was ordered for each site by Census geographic variables to assure geographic dispersion of the sample across the target area. The number of FSUs selected was 250 for Washington and 100 for Denver. Commercial listings of names, addresses, and telephone numbers were then purchased for each sampled FSU.

However, there were no commercial listings for three Washington FSUs that each contained over 300 occupied housing units based on the 1980 Census. There were no listings for these three area segments because they were in an area of Maryland for which Donnelley Corporation did not have telephone directory listings and because vehicle registration records were not available for Maryland. Standard field procedures were used to identify sample housing units for these FSUs or area segments. All screening interviews were conducted in the field for these three segments. Screening interviews were first attempted by telephone for all other area segments.

For all FSUs with commercial listings, the individual listings were the second stage sampling units. A simple random sample of 50 listings was selected without replacement within each Denver segment, and 40 listings were selected within each Washington segment. The Washington sample was later subsampled to approximately 35 listings per segment and some segments contained fewer than the desired number of listings so that the ultimate sample sizes were 4987 sample listings for Denver and 9876 listings for Washington. This design was chosen partly to produce approximately equal sampling weights for the screening sample. Sampling weights for the screenings differ only to the extent that the 1980 Census occupied housing units differ from the number of commercial listings for each FSU.

The Denver and Washington samples of commercial listings were regarded as samples of addresses for the CO study. However, the addresses in the listings had the following problems with respect to locating the sampled housing units. Most residents of apartment complexes usually had identical addresses in the listings; they had the apartment street address in both the telephone book and the commercial listings. In this case, both the name and address fields in the listings were used to locate a specific apartment.

The telephone numbers in the listings were used to obtain screening interviews for all household members at the sample addresses whenever possible. When the telephone number in the listing did not access the listed address, the listing was placed in a pool to be subsampled for field interviews. As seen in Table 1, this occurred for about seven percent of the listings. A telephone number was obtained from the information operator, when possible, for the listings with no telephone number. Telephone numbers were obtained for approximately 10 percent of the listings with no telephone number. The full telephone screening results in Table 1 show that it was not possible to obtain a telephone number for approximately 20 percent of the Denver listings and approximately 10 percent of the Washington listings. This difference occurred because vehicle registration records were available for the entire Denver population but were not available for that portion of the Washington population living in Maryland. This difference is also directly related to the difference in percentage of sample listings that generated complete interviews: 40.1 percent for the Denver sample, and 49.1 percent for Washington.

A subsample of the listings for which a telephone interview was not possible (see Table 1) was selected for field screening. Also, a subsample of listings was selected for a "missed housing unit" (missed HU) check.

The missed HU subsample consisted of 150 listings for Denver and 300 listings for Washington. In each case, the FSUs with the number of 1980 Census occupied housing units 50 percent or more greater than the number of commercial listings were deliberately oversampled as shown in Table 2. The missed HU check was implemented by using standard field listing protocol to produce a unique geographic ordering on Census maps for each FSU. Each listing in the missed HU sample was located in the field. The interviewer then proceeded to the next housing unit as identified by the geographic ordering and checked to see if that housing unit was on the commercial list for the FSU. If not, a screening interview was attempted and the check was continued at the next housing unit. When the next housing unit was found to be on the commercial list, the missed HU check was complete. Technically, a screening interview should not have been conducted if the missed HU was on the complete frame of commercial listings and was simply missclassified with regard to Census block group. Interviews were conducted for all missed HUs, regarding them as not simply missclassified, partly to check the completeness of the listings. In most cases, missed HUs occurred in groups of one or two. In one instance, an entire block face of five HUs was missed (See Table 3). These five missed HUs were regarded as missclassified, and their data were disregarded for analyses and selection of participants for CO monitoring.

The results of the missed housing unit checks for Denver and Washington are summarized in Table 3. For each study site, approximately two percent of the listings were found to not belong to the FSU, or area segment, to which they had been classified by Donnelley Corporation. Although a unique geographic ordering was not possible for listings outside the assigned area segment, a missed HU check was attempted for

these listings. The purpose of this check was mainly to investigate the completeness of the commercial listings. The results in Table 3 for start addresses outside the segment would seem to indicate that clusters of HUs, e.g., block faces, tend to be misclassified occassionally and that random missclassification of individual HUs also occurrs.

When the missed HU start address was inside an apartment complex, implementation of the missed HU check was sometimes quite difficult. Table 3 shows that it was not possible to complete the missed HU check in apartment complexes for about 15 percent of the listings selected for the missed HU check. Since the listings did not generally include apartment numbers, it was necessary to get apartment numbers from mail boxes, apartment managers, or apartment residents. Sometimes these sources proved fruitless. Many of the instances in which a missed HU check could not be begun occurred in restricted-access apartments. Missed HUs seemed to occur more frequently in apartment complexes than in other areas when the check could be implemented. This may be due to the more transient nature of apartment dwellers. Only one missed HU check identified an entire block face (of five HUs) that had been missed by the commercial listings within the selected area segments. The general impression was that the commercial listings provided a reasonably complete listing of housing units.

4. Lousiaina's Social Service Needs Survey

The household residents of the State of Lousiana were the population of inferential interest for the social service needs survey. The sample design was a stratified one-stage, two phase cluster sample. Sampling units were defined as noncompact clusters of housing units constructed within geographic area segments consisting of block groups, enumeration districts, or combinations thereof. The sampling units, or clusters, were constructed to contain an average of 12.52 housing units based upon the 1980 Census data.

The sampling frame consisted of 120,050 unique clusters classified into 12 strata. A sample of 100 clusters was allocated to the strata in proportion to the 1980 population of each stratum. Within each stratum the sample was selected with equal probability and without replacement.

Under phase one of the design, the households contained in each sample cluster were identified using the total set of commercial listings for the area segments containing at least one sample cluster. Both telephone directory listings and vehicle registration records were used by Donnelley Corporation in compiling the listings for Louisiana. During phase one, telephone interviews were attempted for all sample listings using the protocol shown in Figure 1. This protocol made full use of the name, address, and telephone number fields in the commercial listings. The sample household(s) corresponding to a listing were identified by the listed telephone number if it was a working, residential number. Otherwise, the sample household(s) were identified by the address in the listing. The final results from the telephone attempts in phase one are shown in Table 4.

Phase two of the design consisted of obtaining field listings and in-person interviews for a subsample of 53 clusters. There were 23 sample clusters with no commercial listings. These clusters were all included in the phase two subsample. Standard field procedures were used to list the housing units in these 23 clusters. In addition, a subsample of 30 of the remaining clusters was selected for field verification. For each sample listing (commercial listing or field HU listing), an in-person interview was conducted (a) for the household(s) associated with the sample listing, if not already interviewed by telephone, and (b) for the missed housing units linked to each sample listing.

The missed HU procedure for Louisiana differed from that described for the CO study in some ways. Both studies determined a unique geographic ordering for the housing units in each area segment. The missed HU procedure for sample listings inside the area segment was identical to that for the CO study, except that the Louisiana study assumed that housing units with at least one currently listed telephone number and/or vehicle registration were on the commercial lists and simply missclassified. No interview was conducted for these HUs. Moreover, no missed HU check was performed for listings outside the area segment. These HUs were assumed to be covered by the commercial listings for area segments not selected into the sample. As a result, almost no missed HU interviews were performed for Louisiana.

Some problems experienced with the commercial listings for Louisiana appeared to be peculiar to rural areas. In rural areas, the listed addresses were often only route numbers or road names. This made location of the sample households somewhat difficult. However, use of local people as interviewers seemed to alleviate this problem.

Unbiased estimators of population parameters defined by linear statistics are available for this design in the form of difference estimators (See, e.g. Konijn [1973]). In this context, the subsample information can be thought of as providing estimates of the biases affecting the commercial listing sample, e.g., the incomplete frame bias. The bias estimates are then used to correct the list sample estimates.

5. Conclusions

It appears the commercially available listings of residential names, addresses and telephone numbers classified according to Census geographic variables have potential application in sample survey designs. The use of these listings cannot simply be supplemented with interviews for listings without telephone numbers and missed housing units for several reasons. First, listings for apartment complexes and rural roads tend to all have the same address. This makes field location of sample housing units very difficult. Second, missed housing units cannot generally be identified reliably. Hence, the dual frame design discussed in Section 2 is recommended. The commercial listings are used for telephone interviews based upon the telephone directory listings only. An independent field sample is used

to compensate for the undercoverage bias associated with the telephone directory listings. This procedure can result in reduced survey costs since field listings of housing units and field interviews are required for only half of the sampled area segments.

REFERENCES

1. Casady, Robert J. and Sirken, Monroe G. [1980]. A Multiplicity Estimator for Multiple Frame Sampling. Proceedings of the American Statistical Association Section on Survey Research Methods, 601-605.
2. Chromy, James R. [1979]. Sequential Sample Selection Methods. Proceedings of the American Statistical Association Section on Survey Research Methods, 401-406.
3. Konijn, H.S. [1973]. Statistical Theory of Sample Survey Design and Analysis. American Elsevier Publishing Company, New York, 126-132.

INTERVIEW FLOW CHART

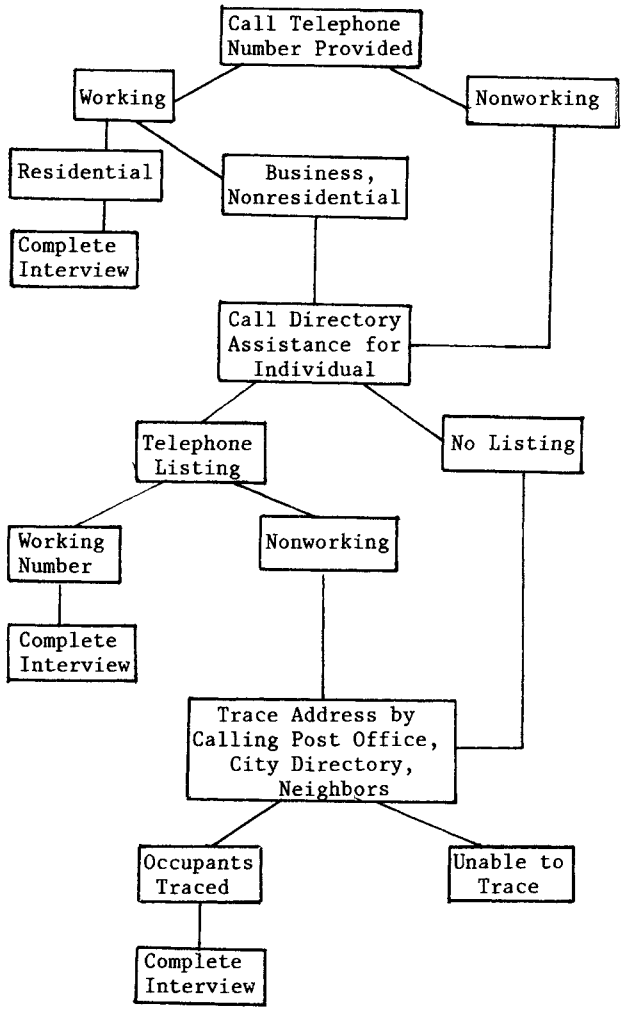

Figure 1. Protocol for the Telephone Phase of the Louisiana Social Service Needs Survey

Table 1. Final Telephone Results for the Sample of Commercial Listings for the EPA's CO Study

Final Result	Denver		Washington	
	N	%	N	%
Completed Interview	1997	40.1	4245	49.1
Partial Interview	13	0.3	32	0.4
Refusal	604	12.1	766	8.9
No Reliable Respondent	27	0.5	46	0.5
Nonresidential Number	91	1.8	187	2.2
Wrong Address[1,5]	350	7.0	609	7.0
Entire Household Moving[2]	92	1.8	187	2.2
Ring No Answer	207	4.2	503	5.8
Final Busy	4	0.1	4	0.1
Nonworking Number[5]	436	8.7	997	11.5
No Phone Number[3,5]	1032	20.7	782	9.0
Final Phone Problem[5]	7	0.1	8	0.1
Other[4]	127	2.6	277	3.2
Total	4987	100.0	8643	100.3

[1]The sample was regarded as an address sample. When the listed telephone number did not match the listed address the listings was placed in a pool to be subsampled for field interviews.

[2]When the entire household was planning to move no screening interview was attempted. There was no reason to screen individuals who would not be available for personal CO monitoring.

[3]No Phone Number was the final result for sampled Donnelley listings with no telephone number when a telephone number could not be obtained from the information operator. A telephone number was obtained from the information operator for about 10 percent of the Donnelley listings with no telephone number.

[4]Includes answering machines, etc.

[5]A subsample of the addresses in these categories was selected for field screening.

Table 2. Stratification of FSUs for the Missed HU Sample

Site	Stratum Number	Stratum Definition	Total FSUs	Sample FSUs
Denver	1	$0<EHUS^{1}\leq 1.5$	92	13
Denver	2	$EHUS > \bar{1}.5$	7	2
Washington	1	$0<EHUS\leq 1.5$	204	24
Washington	2	$1.5<EHUS\leq 2.5$	34	4
Washington	3	$EHUS > \bar{2}.5$	9	2

[1]EHUS is defined to be the ratio of the number of 1980 Census occupied housing units for the FSU divided by the number of Donnelley listings for the FSU.

Table 3. Results of Missed HU Checks

Result	Denver[1] N/%	Washington[2] N/%
A. Start Address Inside Area Segment		
1. Completed missed HU check and found no missed HUs.	108/72.0	203/67.7
2. Completed missed HU check and found one or more missed HUs	18/12.0[3,4]	22/7.3
3. Invalid start address	8/5.3	2/0.7
4. Could not locate start address due to incomplete Donnelley listing	0/0.0	1/0.3
5. Could not identify the apartment at which to begin the missed HU check	12/8.0	39/13.0
6. Found one or more missed HUs but not able to complete missed HU check (unable to match names in the Donnelley listings to apartment numbers)	1/0.7	4/1.3
7. Start address and next address were both office complexes; missed HU check was aborted	0/0.0	8/2.7
8. Start address was inside an old age or convalescent home; missed HU check aborted as group quarters were ineligible	0/0.0	16/5.3
B. Start Address Outside Area Segment		
1. Completed missed HU check and found no missed HUs	0/0.0	2/0.7
2. Completed missed HU check and found exactly one missed HU	0/0.0	1/0.3
3. Could not identify apartment at which to begin the missed HU check	1/0.7	0/0.0
4. Aborted missed HU check after traveling one mile, to first corner, or listing nine missed HUs	2/1.3[4]	2/0.7[4]
Total	150/100.0	300/100.0

[1]Field work done by Research Triangle Institute.

[2]Field work done by PEDCo Environmental, Inc. under a separate contract.

[3]In one case an entire block face of five HUs was missed. The data for these five HUs was disregarded. These missed HUs were regarded as misclassified.

[4]The data for these missed HUs were disregarded. These missed HUs were regarded as misclassified.

Table 4. Final Telephone Phase Results for the Sample of Commercial Listings for the Louisiana Social Service Needs Survey

Final Result	N	%
A. Complete Interview	591	65.67
B. Breakoff/Partial Data	23	2.56
C. Vacant	10	1.11
D. Demolished/Merged/Not HU	14	1.56
E. Vacation/Second Home	2	0.22
F. Disconnected/Non-Working Number	6	0.67
G. No Phone Number	14	1.56
H. No Answer/Busy/Not at Home	15	1.67
I. Language Barrier	1	0.11
J. Not Available During Survey	2	0.22
K. Mentally Incompetent/ Physically Disabled	4	0.44
L. Refusal	114	12.67
M. Unable to Trace/Locate	75	8.33
N. Nonpublished Number	29	3.22
Total	900	100.01