IDENTIFYING GEOGRAPHIC LOCATION FROM RESPONDENT INTERVIEWS USING RDD SURVEYS

David A. Marker, Joseph Waksberg, and Leslie Athey, Westat, Inc.
1650 Research Boulevard, Rockville, MD 20850

KEY WORDS: Random digit dialing, oversampling, telephone surveys

Introduction

It is sometimes necessary to identify a respondent's geographic location when using random digit dialing (RDD) surveys. For example it may be necessary to identify who lives in a SMSA versus non-SMSA, urban versus rural area, or over- versus under-sampled parts of a county or city. In an area probability survey identification of these characteristics is usually made from the maps used to define and select the area segments, but in an RDD survey we must rely on the respondents' ability to describe their locations. Obtaining information on whether a residence is within a specific city or county limits is fairly simple. It is more difficult to deal with other kinds of geographic boundaries.

This paper discusses the results of the RDD screener survey that was part of the National Survey of Pesticides in Drinking Water Wells (NPS) conducted for the U.S. Environmental Protection Agency. Geographic location was necessary to satisfy two survey requirements. First, to be eligible for this survey the respondent had to live in rural parts of specified sampled counties. Second, areas with ground water very vulnerable to pesticide contamination (near the surface, sandy soil, or with certain other characteristics) and densely farmed were targeted to be oversampled. It was therefore necessary to identify a respondent's county, urban/rural status, and whether he or she lived in a targeted part of the county. The highly vulnerable areas are not shown on standard maps that respondents might have access to, and therefore had to be described by the telephone interviewer. Similarly, urbanized area boundaries are not shown on maps to which respondents will have access, and respondents' location with respect to those boundaries are necessary to determine urban-rural residence.

This paper contains the methods used to describe the target areas to the respondents, discusses how successful we were at identifying the location of respondents (e.g. rural, oversampled), and examines the types of questions which respondents seemed to have most trouble answering.

National Pesticide Survey Screener Questions

Although the National Pesticide Survey had two parts - a Community Water System (CWS) survey, and a Domestic Well Survey (DWS) - the geographic identification issues only occurred in the Domestic Well Survey. This is essentially a survey of rural households using wells for their water supply. Unlike most surveys using RDD for sample selection, a multi-stage sample was used. Counties were used as primary sampling units since it was necessary for interviewers to visit the household to take samples of well water. The sampled wells were identified by using a RDD screener questionnaire in 84 counties. Since the urban and target boundaries varied from county to county these results can be viewed as coming from 84 separate small scale surveys.

Telephone exchanges do not respect county boundaries, and the first step was to identify the telephone exchanges which overlapped the sample counties, select an RDD sample and screen out residences outside the sample counties. Urban areas of 2,500 or more people and urbanized areas defined in the 1980 Census, were excluded from the survey. When these areas corresponded to city limits the respondents were simply asked if they resided within the city limits of cities or towns in the sampled counties, with each place named by the interviewer. In counties containing urbanized areas, we had to ask a series of questions which contained descriptions of the perimeter of the urbanized area. In the time since the 1980 census it is likely that urban areas have expanded; thus when the area defined as "urban" did not have a boundary that could be identified by most respondents, the description was extended to the next identifiable boundary. We thus have urban, extended urban, and rural areas of a county.

A similar procedure was followed for areas targeted for oversampling. A series of questions were asked describing the boundaries of the areas to be oversampled. These boundaries typically did not correspond to roads or other easily describable boundaries; it was therefore necessary to use those descriptive boundaries that most closely fit the boundaries of the target area. In addition some target areas were not oversampled because they were so small as to be almost impossible to describe via the telephone. (It is unlikely that these areas would be oversampled in an area probability survey either, since they were typically small parts of EDs or block groups which cannot be identified in Census records.) This resulted in four categories of geography with respect to targeted areas: target, extended target, undefinable target, and not target. Figures 1 and 2 provide examples of part of a county map with targeted areas and the corresponding set of questions, respectively.

It is commonly assumed (and not tested in this study) that the general population in the United States cannot describe locations with regards to north, south, east, and west. It is therefore not appropriate to use such terms in describing boundaries. Thus the questions used in the RDD screener were of two types: "Do you live within x miles of Route 37?", or "Do you live between US 25 and Badlands State Forest?" One of the main purposes of this paper is to compare the relative success of these two types of questions.

Before presenting our results it is important to include two cautionary notes. First, the accuracy found in the NPS is partially a result of the well-trained interviewers who were able to work with the respondents by reviewing maps of the sampled counties while conducting the interviews. Second, the ability to determine boundaries that would be known to respondents depended upon the elaborate procedures undertaken (verifying information with local post offices, etc.) by the NPS mapping staff.

Results

As part of the National Pesticide Survey we screened over 13,500 households. In this paper we examine three types of results from the respondents to the 84 different RDD screeners who were included in the field survey. The RDD computer assisted telephone interviews (CATI) screeners selected 835 eligible wells for the DWS from the rural parts of the 84 counties. Responses were obtained and locations recorded by the field interviewers for 719 of these wells. By comparing these actual locations (we are assuming no mapping error by the field interviewers) with the classifications from the telephone screener, we can answer the following questions.

1) How frequently did people mis-identify their county of residence?

2) How frequently did people who lived in urban areas get identified over the telephone as living in rural areas?

3) How frequently did over- and under-sampled areas get confused?

We can only measure county mis-identification in one direction, respondents who stated they lived in a sampled county when they actually didn't. However, since the only telephone prefix numbers that were used had some telephone numbers in the sample counties (and many prefixes only contained numbers in the sampled county), the number who made the opposite mistake (claiming to live in the sampled county when they actually don't) should be very small. Mis-identification of county is likely to be a highly regional phenomenon. In the New England states most political and economic activity is either at the state or town level, with little activity at the county level. It is therefore much more likely for these residents to be confused over their county of residence.

To be eligible to be selected for the NPS one had to live in a rural part of the county. Thus all of the 719 respondents were identified by the RDD survey as being rural, with none urban (or extended urban). While those living in the extended urban area still meet the eligibility requirements for the survey, their inclusion in the sample would indicate a mis-classification by the telephone screener since there is no way to differentiate them from urban households.

Similarly the screener only differentiated between oversampled areas (target or extended target) and undersampled areas (non-target or undefinable target). Mis-classification errors occur when someone in the target (or extended target) area is not oversampled, or if someone is in the non-target (or undefinable target) area and is oversampled. How closely the areas that the screening survey oversampled matched the actual target areas is a related question of interest, but is not an indication of how successfully geographic locations can be identified using a telephone screener.

Of the 835 wells selected for the NPS after being identified on the CATI screener as being in the rural parts of the 84 sampled counties, 6 (less than 1%) turned out to be located in another county. Four of these 6 wells were in Rhode Island, where county boundaries appear to be of little importance to residents.

Of the 719 sampled and plotted wells only 1 (less than 1 percent) was found to actually be located within an urban area. Six others were in extended urban areas which, although eligible for the survey, should not have been included in the sample if the telephone questions had been answered correctly. Of the seven wells misclassified as rural, two supplied the wrong ZIP code and thus were never asked the city boundary questions. One incorrectly responded to a question about living between a road and a river, and the other four were right on the boundary. Thus over 99 percent of all respondents identified by the telephone screener as located in rural areas were indeed rural.

Table 1 shows how well the telephone screener's decision on whether or not to oversample each of the 719 sampled and plotted wells matched the target status of the well as identified by the field interviewer. 353 of the 414 wells (85 percent) that were oversampled really did deserve it based upon being in either actual or extended target areas. Similarly, 273 of the 305 wells (90 percent) that were not oversampled were correctly in either the non-target or undefinable target areas.

Of the 719 wells with known locations, 93 (13%) had discrepancies on their target status. Four of these were also among the 7 wells which had urban/rural discrepancies, resulting in a total of 96 cases with differences between the CATI and field locations. In sixteen of the cases, either the boundary lines were not

clear or how the case was treated by the CATI program (e.g., when the respondent answered "don't know") was not clear. Thus there are 80 cases with clear discrepancies.

In each of the 84 county screener questionnaires immediately following asking the respondent's county, the respondent was asked their ZIP code. Based on the response to this question the CATI program determined the appropriate set of questions to ask. This served to minimize the number of location questions asked of any respondent.

In 10 of the 719 cases (1%) the respondent gave an incorrect ZIP code to the CATI interviewer. The other 70 discrepancies are shown in Table 2. Ideally the number of discrepancies for each type of question would be compared to the number of times that type was asked. Due to the multitude of skip patterns each CATI interview could follow, we have used an approximate methodology. The third column of Table 2 shows the number of times that type of question appears on any of the 84 screener surveys. Assuming each skip pattern is followed an equal number of times, this methodology will provide an accurate measure of the relative frequency of errors for the different types of questions.

While the overall discrepancy rates are small, there are certain types of questions that seem to cause more problems than others. "Do you live between a road and [another landmark]" had 50% more discrepancies per 100 occurrences on a questionnaire than did "Do you live within x miles of [a landmark]", 3.95 to 2.66, respectively.

For both types of questions, by far the most troublesome type of landmark was a state. On the other end of the spectrum, roads seem relatively easy to identify.

Summary

Respondents appear quite able to identify their location to telephone interviewers in the great majority of cases. Over 99% of those identified as rural were indeed rural. In 87% of the cases the respondent was correctly placed into oversampled/non-oversampled areas. While respondents had most trouble providing location relative to state boundaries, they had little trouble when asked about roads in their area. We note that the errors in urban-rural classification can introduce potential bias in the survey results where errors in the oversampled/undersampled areas contribute only to variance. Luckily, the urban-rural errors were negligible. The 13 percent of these errors should not have an important effect on the variances. However, oversampling rates should not be set too high when this type of screening is used because even an error rate as low as 13 percent can add significantly to the variances if there is a very wide discrepancy in the sampling rates.

These results are reasonably encouraging for telephone surveys. It is possible to oversample and/or define eligibility of respondents based on their geographic location although it takes a lot of careful work to determine which boundaries both describe the areas in question and are identifiable to respondents.

Acknowledgements

The authors would like to acknowledge the support of the U.S. Environmental Protection Agency and in particular the EPA staff of the National Pesticide Survey. We also thank ICF Inc. who worked with us on this contract. The work upon which this publication is based was performed pursuant to Contract Number 68-D0-0006 with the Environmental Protection Agency.

Figure 2. Example of questions used to identify target/ non-target areas.

If the respondent's ZIP code is **68045**, the following questions will be asked at **B4:**

B4a. Do you live between Highway 77 and Cuming County?
YES Go to B4b
NO Skip to B4c

B4b. Do you live within $\frac{1}{2}$ mile of Cuming County?
YES Skip to B5b or C0
NO Skip to C0

B4c. Do you live within $\frac{1}{2}$ mile of Bell Creek? (Make sure respondent understands you are asking about the creek and not the township.)
YES Skip to C0
NO Go to B5b or C0

If the respondent's ZIP code is **68061**, the following questions will be asked at **B4:**

B4a. Do you live between Highway 75 and (the state of) Iowa?
YES Skip to C0
NO Go to B4b

B4b. Do you live between Route 32 and Washington County?
YES Go to B4c
NO Skip to B5b or C0

B4c. Do you live within $1\frac{1}{2}$ miles of Highway 75?
YES Skip to C0
NO Go to B5b or C0

Table 1. Field vs. CATI target determination

Oversampling	Field determination	CATI determination: Over-sampled	CATI determination: Not over-sampled	Total
Desired:				
	Actual target	**315**	**25**	**340**
		92.6	7.4	100.0
		76.1	8.2	47.3
	Extended target	**38**	**7**	**45**
		84.4	15.6	100.0
		9.2	2.3	6.3
Not desired:				
	Non-target	**59**	**272**	**331**
		17.9	82.1	100.0
		14.3	89.2	46.0
	Undefinable target	**2**	**1**	**3**
		66.7	33.3	100.0
		0.5	0.3	0.4
Total		**414**	**305**	**719**
		57.6	42.4	100.0
		100.0	100.0	100.0

Table 2. Frequency of discrepancies by type of question

Question type	Number of discrepancies	Number of occurrences on questionnaires	Discrepancies per 100 occurrences
Between road and state	3	40	7.50
Between road and county	14	294	4.76
Between road and river	6	140	4.29
Between road and town	1	25	4.00
Between road and lake	2	51	3.92
Between road and road	15	489	3.07
	41	1039	3.95
x miles from state	5	31	16.13
x miles from lake	3	55	5.45
x miles from river	10	283	3.53
x miles from road	7	516	1.36
x miles from county	1	92	1.09
	26	977	2.66
Between river and county	1	43	2.33
City limits	2	297	.67
Total	70	2356	2.97

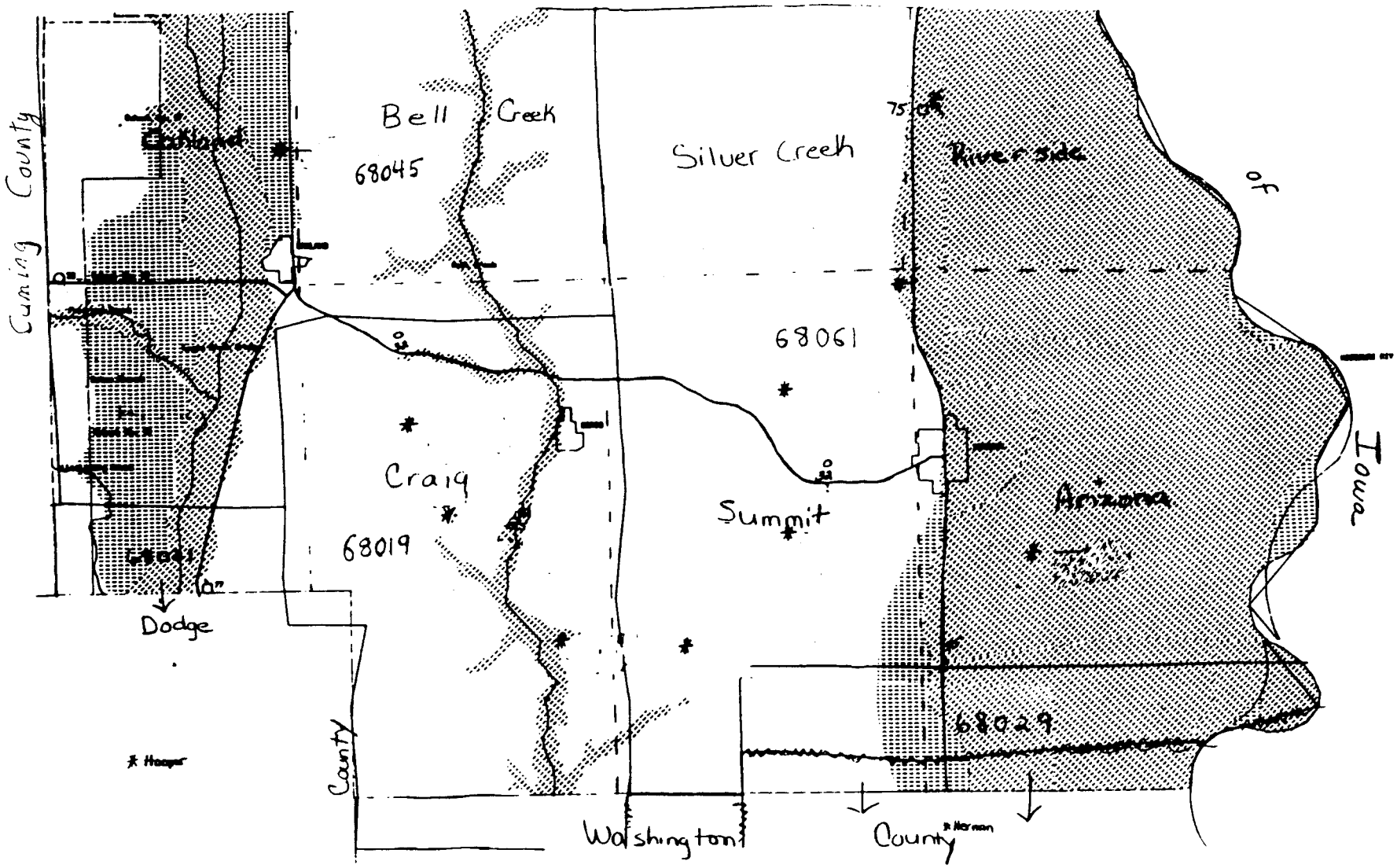

Figure 2. Partial map of county, with shaded target areas