EXAMINING THE RELIABILITY OF SELF-REPORTED LOCATION INFORMATION IN RDD TELEPHONE SURVEYS

Douglas Willson, Greg Mahnke, Julie Bouffard, and Geoff Halsted
Macro International Inc.
Douglas Willson, Macro International Inc., 126 College Street, Burlington, VT 05401

Key Words: RDD Surveys; Geographic Eligibility; Bias; Undercoverage.

1. Introduction

Sampling frames for RDD telephone surveys that target households within relatively small geographic areas such as towns or counties are typically constructed using the exchanges that serve households within these areas. Unfortunately, telephone exchange boundaries rarely coincide exactly with the physical geographic boundaries of interest for these surveys. In these instances, sampling frames with high or complete coverage of the households in the target area will include households who reside outside the area. In order to prevent geographically ineligible respondents from biasing survey estimates, respondents are asked to report on their eligibility for the survey during the interview, and only eligible households are included in the final data set. This paper examines the reliability of this type of self-reported location information using data from a large-scale telephone survey.

For small geographic areas, the proportion of geographically ineligible telephone households included in an RDD sampling frame can be substantial, and the overall cost of the survey can be significantly increased because of the costs of screening ineligibles. Even in situations where the proportion of ineligibles is relatively small, responses to location questions must be complete and accurate to prevent the introduction of coverage biases. Some authors have suggested the use of dual frame survey designs to reduce costs and coverage biases for surveys of small geographic areas (Schejbal and Lavrakas, 1994). Research investigating the quality of self-reported location information is essential for evaluating the choices between different sampling frames, survey designs, and questionnaires.

This paper compares self-reported location information from RDD survey respondents who reside in directory listed households, with geocoded location information for the same households from a national telephone directory. We investigate a variety of general hypotheses concerning how individuals reference and interviewers record location information, and how this information relates to sampling frame characteristics, questionnaire design, and the characteristics of the physical geography of interest. We consider survey responses from the FY97 Area-Specific Section 8 Fair Market Rent (FMR) Surveys conducted for the Department of Housing and Urban Development by Macro International during the first four months of 1997. The geographic areas of interest in these surveys are FMR areas which are defined (except in New England) by a county or collection of counties; in New England, FMR areas are defined by collections of towns or cities and the appropriate boundaries are township boundaries. Survey information is collected from 1 and 2-bedroom renters in each area that is surveyed.

More specifically, we investigate the potential for bias in survey estimates resulting from measurement errors in self-reported location information. Errors in location information can be introduced from a variety of different sources. Interviewers may record the information incorrectly, through keystroke errors or through misinterpretation of responses. Respondents also refer to their place of residence in many different ways. For example, some individuals identify their town of residence with the name of their local post office, yet ZIP code boundaries do not necessarily correspond with official town boundaries. Nonresponse is another problem. Respondents in urban areas are much less likely to know their county of residence than respondents in rural areas, and are also much more likely to refuse to answer questions about the location of their residence.

We would like to thank Alan Fox for providing helpful comments on an earlier draft, and Amy Starer from Genesys Sampling Systems and Sally Ann Ciarlo from Donnelley Marketing for providing the census tract geocoding used in this study. The views expressed in this paper are those of the authors and not necessarily those of the Department of Housing and Urban Development, Genesys Sampling Systems, or Donnelley Marketing.

We consider two types of measurement error models for categorical data that differ according to whether or not the matching procedure for geocoding is assumed to be error-free. Record-check studies often assume that measurement errors are confined to the survey measurements, yet it is doubtful that telephone directory listings are error-free, or that geocoding is an error-free process. As a result, it is important to allow for the possibility that the geocoded records are measured with error as well.

We also investigate whether the geographic complexity of the target area, as proxied by the number of towns or counties in the FMR area being surveyed, adversely influences the degree of recorded ineligibility. For the FMR surveys, the location question included a disk-based look-up within the CATI program to reduce open-ended responses and interviewer recording error. For each FMR area, a list was displayed containing the counties (or towns) that comprised the FMR area, as well as contiguous counties (or towns) that were outside the area. Interviewers were required to select a name from the hard coded list, or to record an open-ended response for other self-reported county (or town) names that were not on the list. As the number of counties or towns displayed on the interviewing screen increases, it may be the case that the difficulty of the interviewer's location coding task increases, possibly affecting recorded location information.

It is important to realize that there are a variety of difficulties in generalizing the results in this paper to household surveys or surveys of other target populations. Perhaps most important, the FMR survey targets rental households, and the analysis presented below only focuses on reported location information for listed rental households. Listed rental households may differ substantially from unlisted rental households, as well as listed and unlisted households with other dwelling types. Telephone listings for rental households may also be much less complete than for other households, either because directory updates occur at fixed intervals and renters move more frequently relative to the population in general, or simply because renters are less likely to list their telephone number. We present evidence to support both of these hypotheses below. One additional difficulty is that the results presented here apply to FMR areas, not specifically to towns, counties, or zip-codes. Although respondents report on their town or county of residence, self-reported and geocoded locations are recoded as in or outside the FMR area of interest, because this is the most important classification for the survey in question.

The paper will proceed as follows: Section 2 provides a brief description of the data sources employed in the paper. Section 3 presents the empirical results. Section 4 concludes and suggests avenues for future research.

2. Data Sources

The research presented in this paper is based on two data sources: 1) The FY97 Section 8 Area-Specific Fair Market Rent Telephone Surveys; and 2) The DQI^2 Consumer File from Donnelley Marketing Inc. This section of the paper describes each one of these sources.

2.1 FY97 Area-Specific FMR Surveys

Section 8 Fair Market Rents (FMRs) form the basis for housing subsidy payment levels under the Certificate and Voucher programs operated by the Department of Housing and Urban Development (HUD). HUD establishes FMRs by geographic area; in general, they are intended to represent the market cost of privately-owned, decent, safe, and sanitary rental housing. In order to ensure that FMRs accurately reflect rental costs in different housing markets, HUD conducts Area-Specific Surveys to produce a point estimate of the 40th percentile rent level in relatively small geographic areas such as towns, counties or collections of counties. These estimates are used to re-benchmark the FMR for a single FMR area. HUD also conducts regional surveys which are designed to produce estimates of annual changes in gross rents, called annual adjustment factors (AAFs), in broad geographic areas defined by the metropolitan and non-metropolitan portions of each of the ten HUD regions. These estimates are used to re-benchmark FMRs in areas that are not already covered by annual Consumer Price Index surveys or a current Area-Specific Survey. This paper examines survey responses from 52 Area-Specific surveys conducted in 1997.

The sample for these surveys was drawn from the list-assisted sampling frame developed by Genesys Sampling Systems. Hundred banks with at least two listed residential households were included in the frames. For each survey, the census tracts comprising the geographic area were selected, an exchange report documenting coverage and noncoverage for each exchange serving these tracts was generated, and the exchanges predominantly serving the area were selected for the frame. In general, excluding eligible households (undercoverage) was viewed as a much more serious problem than including ineligible households; including ineligibles increases the

cost of the survey but does not bias survey results, if the location screening questions obtain reliable location information. As a result, some exchanges that were primarily outside the area of interest were also included in area frames in some instances, to ensure sufficient coverage. For the FY97 Area-Specific Survey areas, coverage for listed households averaged over 99 percent, and noncoverage ranged from slightly above zero to a maximum of 22.6 percent, with an average noncoverage across areas of 3.6 percent.

At the end of the interview, FY97 FMR Area-Specific Survey respondents outside of New England were asked the following location question:

In what county is your housing unit located?

For FMR areas within New England, a similar question was asked regarding town or city of residence, and eligible and contiguous but ineligible town or city names were included in the displayed list. For respondents outside New England that did not know their county name, a town name question was asked. All respondents were also asked to report their ZIP code.

2.2 The DQI^2 Consumer File

The DQI^2 Consumer File from First Data InfoSource/Donnelley Marketing was the source for the geocoded location information that is used in the analysis below. Telephone directories are one of nine primary data sources in the DQI^2 file. Telephone information is captured from over 4,500 telephone directories annually, using both scanning and direct data entry technologes. After the telephone directory information is captured, the records are matched against InfoSource/Donnelley's Geographic Base File (GBF). The GBF is compiled from USPS products which list streets, corresponding block ranges, and ZIP codes. The GBF is used to append Census state, county, tract and block group codes. Records that do not match the GBF on a street level are assigned a pseudo Census geocode, and an extra attempt to match these records using monthly postal datbase updates is performed.

Genesys Sampling Systems performed the actual geocoding for this study. All telephone numbers associated with completed interviews from the FY97 Area Specific FMR surveys were passed against the DQI^2 Consumer File. Each record that matched (based on the 10-digit phone number) was coded with the DQI^2 geocodes for FIPS state/county, Census tract, ZIP code, and OSLO indicator. The OSLO indicator noted records that were pseudo geocodes without a complete address. For the analysis presented in this paper, OSLO records were not included.

3. Empirical Results

3.1 FY97 Area-Specific Survey Results

Table 1 presents unweighted cell counts and frequencies that describe the relationship between self-reported location information and geocoded location information for the listed households from the FY97 Area-Specific surveys. Location information is coded as "Eligible" or "Ineligible", depending on whether or not the respondent was in the FMR area of interest, although the actual survey question referred to county or town of residence. Of the 27,331 individuals who responded to the surveys, 11,046 (40 percent) were listed households with a valid geocoded census tract. This number is much lower than listed rates for households in general, because renters move more frequently and published directory listings will not include individuals who have recently moved into an area. (The corresponding rate for zip code matches is higher (45 percent), illustrating the additional degree of difficulty associated with geocoding to the tract level. However, this rate is still dwarfed by the national listed rate of approximately 65 percent.) Also, renters are less likey to list their telephone number, relative to other households. While individuals who have moved within the last 15 months represented approximately 42 percent of the respondents, valid geocoded census tracts were obtained for only 26 percent of these households. Of the remaining 58 percent who had resided in their current unit for more than 15 months, 51 percent had valid geocoded census tracts.

Diagonal elements in Table 1 represent situations where the self-report and geocode agree, *i..e* individuals who say they live outside the target area are geocoded as outside, and vice versa. Approximately 98.9 percent of the self-reported location responses match the geocoded responses. For the remaining 123 cases (1.1 percent), 63 self-report as ineligible and are geocoded as eligible, while 60 report that they are eligible while they are geocoded as ineligible. Tables 2 and 3 present similar results for recent movers and stayers respectively. In general, these tables indicate that the pattern of errors is not affected by whether the respondent has recently moved, although the effectiveness of geocoding is substantially diminished for the recent mover population.

3.2 Modeling Measurement Error

Under the assumption that the geocoding process is error-free, several statistics can be used to quantify the magnitude and the variability of the measurement errors. Traditionally, researchers consider *bias*, the error that would be constant over conceptual repetitions of the survey, and *simple response variance*, the variability in survey estimates over conceptual repetitions of the survey, as important quantities to measure. For the analysis presented below, we have weighted the data to account for differential sampling rates across FMR areas.

Define the true proportion of eligble renters with listed telephone numbers that reside in the FMR area is p. If we define θ as the proportion of individuals who are actually eligible but who will self-report as ineligible, and define ϕ as the proportion of individuals who are ineligble but who will self-report as eligible, sample proportions from weighted versions of Tables 1-3 can be used to estimate the probabilities and associated parameters in the classification matrix given in Table 4.

If p_1 is the sample proportion of individuals who self-report as eligible, the bias B can be written as (Biemer and Forsman, p. 917):

$$E(p_1) = p + B \tag{1}$$

where

$$B = -p\theta + (1-p)\phi \tag{2}$$

B can obviously estimated using the difference between the off -diagonal sample proportions in the contingency table. For the weighted versions of Tables 1-3, the estimated values for B (0.12, 0.03, 0.27) are all less than one percent, and the null hypothesis that the bias is zero cannot be rejected (at the one percent level) in all cases.

A second quantity that is often analyzed in this literature is the simple response variance (SRV), which captures the variability in survey estimates over conceptual repetitions of the survey. An estimate of the SRV can be calculated as (Bureau of the Census (1985)):

$$\hat{SRV} = \frac{1}{n}\left(\frac{n_{11}n_{21}}{n_{.1}} + \frac{n_{12}n_{22}}{n_{.2}}\right) \tag{3}$$

The estimated values of SRV for the weighted versions of Tables 1-3 are also quite small (0.005,0.003,0.004), and are a direct reflection of small estimated probabilities of misclassification. Biemer and Forsman (1992) point out that this estimate of SRV is biased, but the bias is small relative to the magnitude of SRV.

When both survey measurements and geocodes are measured with error, the bias cannot be estimated directly. In this situation, researches typically concentrate on the gross difference rate g, which is estimated using the sum of the off-diagonal proportions. If the survey measurements and geocodes are assumed to be independent, g measures the sum of the SRV for the survey measurements, and the SRV for the geocode measurements. In the absence of information concerning the relative variability of the measurement errors in each case, it is not possible to determine how much of the gross difference is attributable to measurement error in survey responses, versus measurement error in the geocodes. Kulp (1994) describes the directory update procedure underlying the construction of list-assisted RDD sampling frames, and highlights potential sources of error. Given the size of the national directory, there are millions of records where geographic information is missing, assigned or imputed. In addition, compiling and verifying the directory information is a time consuming process, and the sheer magnitude of the exercise provides numerous opportunities for error. From this perspective, the rates of misclassification presented here represent an upper bound on the magnitude of measurement errors. Overall, absolute error rates for location responses in these surveys are probably quite small.

3.3 Investigating the Causes of Misclassification

In order to investigate the possible causes of misclassification, we estimated a binary logit model for the geocode/self-report cell probabilities, where we define the dependent variable as 0 for records that are similarly classified, and 1 for records that are misclassified. For explanatory variables, we considered three dummy variables and an intercept to capture the effects of MSA/Non-MSA classification and whether or not the area was based on county or township geographies. We also included a variable to represent the number of place names displayed on the CATI screen for each area (NPLACE). (We would like to include a variable for shortest distance to the area boundary in future research, in order to capture the possibility that respondents near the boundary may have a higher probability of being geocoded improperly.)

Table 5 presents the parameter estimates for this equation. The likelihood ratio test for the joint significance of the covariates has a value of 85.185 ($p<0.0001$). In terms of individual parameter estimates, it is clear that the odds of misclassification

are highest for nonmetropolitan FMR areas with township boundaries, and for areas with township boundaries in general. The only variable that is not statistically significant (at the 5 percent level) is NPLACE, which represents the number of place names on the CATI screen for each area. This may suggest that the number of place names on the CATI screen does not significantly affect interviewer coding behavior; alternatively, the effects of this variable may already be captured by the dummy variables representing FMR areas with township boundaries.

3.4 Some Sources of Geocoding Errors

To obtain some descriptive information on possible sources of measurement error for these surveys, we plotted exact geocoded locations for individuals with discrepancies between self-reported and geocoded location information, for one of the FY97 survey areas. The selected survey area contained four separate non-metropolitan counties in Minnesotta. Geocoded households were matched against a CD-ROM white pages directory called SelectPhone from Pro-CD. This database is not as up-to-date as the Donnelley file maintained by Genesys, but contains a latitude and longitude measurement for each listed household. In this area, 230 of 231 previously geocoded listed households were matched, including all 7 discrepancies. These geocoded locations are plotted in Figure 1.

In this survey area, it is clear that the majority of discrepancies are located near the boundaries of the survey area. We conjecture that discrepancies in the center of the FMR areas probably represent interviewer coding or respondent errors, while discrepancies on or near the boundaries probably represent geocoding errors.

4. Conclusions

This paper has examined the differences between self-reported and geocoded location information for a large scale RDD survey of renters. We find that the proportion of listed households where the self-reported location information and the geocoded information disagree is quite small. Once it is recognized that the geocoding process is not 100 percent error free, it appears that the measurement error in location information in these surveys is not a significant source of bias for survey estimates. It is important to remember that these results are not necessarily generalizable to surveys of other populations, or with other geographies. In future research we hope to examine other surveys with different target populations, physical geographies, and questionnaire structures.

5. References

Biemer, P., and Forsman, G., (1992) On the quality of reinterview data with application to the Current Population Survey, *Journal of the American Statistical Association*, 87, pp. 915-923.

Biemer, P., and Trewin, D., (1997) A review of measurement error effects on the analysis of survey data, Ch. 27, pp. 603-632 in *Survey Measurement and Process Quality*, L. Lyberg et. al. (Eds.), New York; John Wiley and Sons.

Kulp, D. (1994) Dynamics of list-assisted random digit dialing (RDD) frame coverage, *1994 Proceedings of the Section on Survey Research Methods*, pp. 11-18, American Statistical Association.

Schejbal, J., and P. Lavarakas (1994) Coverage error and cost issues in small area telephone surveys, *1994 Proceedings of the Section on Survey Research Methods*, pp. 1287-1292, American Statistical Association.

U.S. Bureau of the Census (1985) *Evaluation of Censuses of Population and Housing*, STD-ISP-TR-5, Washington, D.C.: U.S. Government Printing Office.

Table 1: Location Information, All Cases

	Geocode Ineligible	Geocode Eligible	Total
Self Rpt Ineligible	219 (1.98)	63 (0.57)	282 (2.55)
Self Rpt Eligible	60 (0.54)	10,704 (96.90)	10,764 (97.45)
Total	279 (2.53)	10,767 (97.47)	11,046 (100.00)

Table 2: Location Information, Recent Movers

	Geocode Ineligible	Geocode Eligible	Total
Self Rpt Ineligible	51 (1.71)	28 (0.94)	79 (2.66)
Self Rpt Eligible	20 (0.67)	2,875 (96.67)	2,895 (97.34)
Total	71 (2.39)	2,903 (97.61)	2,974 (100.00)

Table 3: Location Information, Stayers

	Geocode Ineligible	Geocode Eligible	Total
Self Rpt Ineligible	168 (2.08)	35 (0.43)	203 (2.51)
Self Rpt Eligible	40 (0.50)	7,829 (96.99)	7,869 (97.49)
Total	208 (2.58)	7,864 (97.42)	8,072 (100.00)

Table 4: Classification Matrix

	Geocode Ineligible	Geocode Eligible
Self Rpt Ineligible	$(1\text{-p})(1\text{-}\phi)$	$p\theta$
Self Rpt Eligible	$(1\text{-p})\phi$	$p(1\text{-}\theta)$

Table 5: Parameter Estimates

	Parameter	S.E.
Intercept	-4.525	0.2197
Metro/County	-0.546	0.2412
Metro/Town	1.293	0.3245
Non-Metro/Town	1.891	0.3857
NPLACE	-0.006	0.0163

Figure 1: Blue Earth, Steele, Rice, and Waseca Counties, MN Pooled FMR Area
FY97 Area Specific FMR Survey
Discrepancies Between Self-Reported and Geocoded Location