On the Quality of Ancillary Data Available for Address-Based Sampling

Charles DiSogra,[1]J. Michael Dennis, [1] Mansour Fahimi[2]
[1]Knowledge Networks, 1350 Willow Road, Menlo Park, CA 94025
[2]Marketing Systems Group, 13416 Bonnie Dale Dr, North Potomac, MD 20878

Abstract

A feature of address-based sampling (ABS) is versatility of the sample frame where many ancillary data can be appended to an address. Commercial databases, e.g., Experian, infoUSA, Acxiom are used to append observed and modeled information at various levels of aggregation. This enables researchers to develop more efficient sample designs and broaden analytical possibilities with expanded sets of covariates. While quality of ancillary data is of concern for researchers, the literature provides only anecdotal assessments on accuracy. Relying on surveys and KnowledgePanel® recruitment samples that employ ABS, the authors present results of comparisons between an array of ancillary data and corresponding observed values collected directly from the responding households. The same ancillary data are also used to demonstrate the ability to analyze non-response bias by comparing the ancillary data available for the invited sample and the subset of recruited study participants.

1. Introduction

Address-based samples (ABS), because they start with a residential address, can be matched to other datasets that we call “ancillary data” in this article. These data have at least two possible uses. We attempt to evaluate these two uses. First, there is the potential to use the ancillary data to draw more efficient, targeted samples. We will refer to this as the “targeting” use (DiSogra, 2010). Second, the ancillary data can be used to analyze non-response at the panel recruitment stage for KnowledgePanel®, as part of a program of research on the topic (Dennis, 2010a).

The address-based sample and ancillary data for the research are provided by Marketing Systems Group (MSG). MSG provided these ABS and ancillary data to Knowledge Networks (KN). KN uses ABS as the sample frame for the recruitment of KnowledgePanel® households.

2. Data Source

The actual sample provided by MSG is derived from the U.S. Postal Service Computerized Delivery Sequence File (CDSF). The CDSF provides approximately 97% coverage of physical addresses. It is frequently updated to include the most recent information on the status of addresses, such as seasonal homes, vacation homes, vacant houses, etc.

Matched to the ABS sample MSG can also provide in many instances the residential telephone number, latitude-longitude location, and ancillary demographic data. At the household level, the following ancillary data are available: Telephone number (landline, match rate 60%+), number of adults, presence of children (yes/no), home ownership (own/rent), and household income (12 categories). At the person level, the following ancillary data are available: Marital status (married/single), education level of head of household, age of head of household, and race/ethnicity (33 categories recoded to four categories).

MSG employs several databases as data sources, including infoUSA, Experian, and Acxiom. These were the data sources used for the ancillary data for our analyses.

Since 2009, Knowledge Networks has recruited panel households for its KnowledgePanel using an address-based sample frame. KnowledgePanel is a probability-based, nationally representative panel of US population age 13 and over (Dennis, 2010b). Sample coverage includes households not having internet access (KN provides laptop computer, free ISP) and Spanish-language dominant households. Because extensive profile data are collected from recruited households through self-report web surveys, the survey data can be compared to the ancillary data made available by MSG for the same households and heads of household.

The data for the analysis of the effectiveness of using ancillary data for targeting is from the address sample fielded in 2008 through early 2010. All sample units fielded for panel recruitment during this time frame are included in the analysis. For most variables, ancillary data for 10,000 or more KN recruited households were available (see Figure 1 for the number of observations per variable and grouping).

The missing data rate for demographic ancillary variables ranges from 5% to 27%. The highest availability rates are for household income and presence of children and lowest for race/ethnicity and educational obtainment.

Some of the demographic ancillary variables are measures of the “head of household.” We expect that there is less accuracy in these “head of household” measures since there are different operational definitions of “head of household” by the respective data sources. There is, in addition, arguably a declining cultural relevance of this term, presenting a further complication in using head of household data for sample targeting.

3. Analysis Method for Evaluating Sample Targeting

The analysis is based on correlations between the demographic ancillary data and the self-reported survey data that Knowledge Networks collected from the panel households as part of the first web survey completed by newly recruited panelists. Also, simple descriptive analyses were conducted; these consist of unweighted frequency distributions from the ancillary and survey data. The analysis was limited to those KN panel households for which ancillary data are available.

The above analyses were completed for all the recruited adult KN panelists (including multiple adults per household), as well as two subsets of the universe of KN panelists:

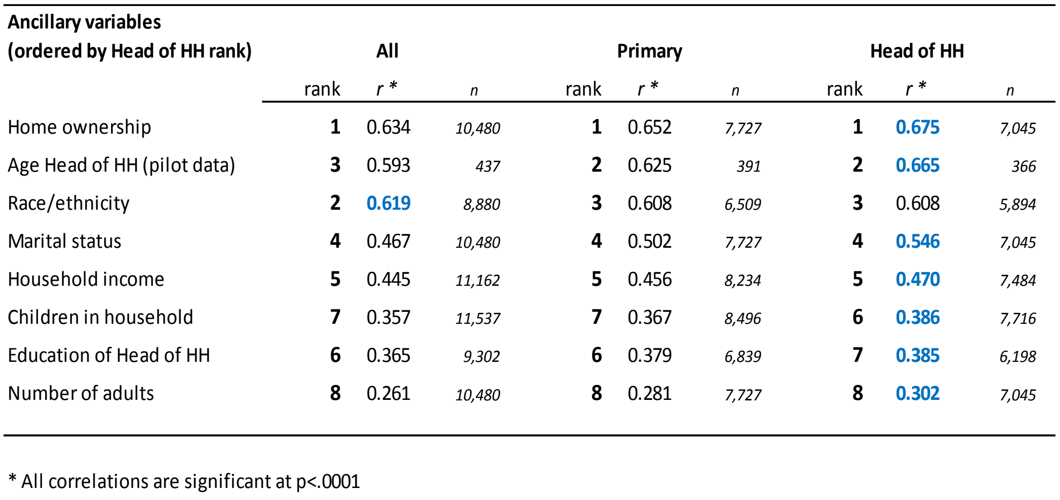
- Primary KN Panel Respondent. This is the first adult listed on the self-administered mail survey recruitment form or entered first online or by telephone during the panel recruitment stage. We refer to this subset as “primary.”

- KN Panel Head of Household. This is the adult in whose name the house or apartment is owned or rented. If there is more than one such panelist in the household, the oldest male is selected as head of HH. We refer to this subset as “Head of HH.”

4. Results from Evaluating Sample Targeting

The ancillary data most closely matched the self-report survey data on home ownership, age of household, and race/ethnicity, ranging from Pearson-R correlations of 0.675 to 0.608 for the subset of head of household panelists. Note, however, that the age of the head of household statistics are based on a small number of observations (n=437 for "all"). Future analyses will include age of head of household for all sample units (less missing data). The lowest match rates were evident for the variables of educational obtainment and the estimate of the number of adults in household.

Figure 1: Correlations between Ancillary and KN Survey Data

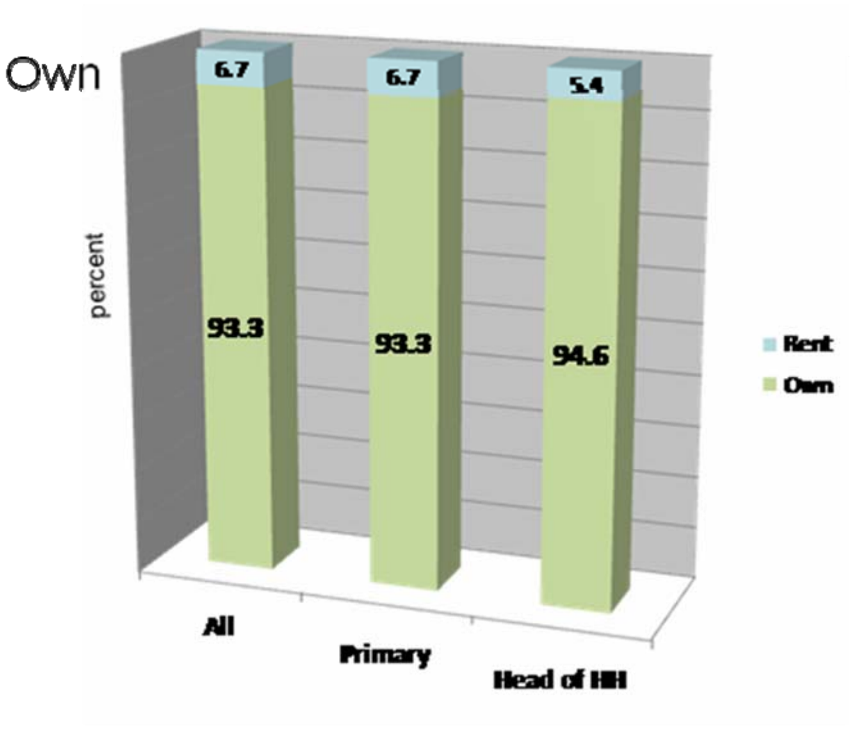


Ancillary variables **(ordered by Head of HH rank)**		**All**			**Primary**			**Head of HH**	
	rank	*r **	*n*	rank	*r **	*n*	rank	*r **	*n*
Home ownership	**1**	0.634	*10,480*	**1**	0.652	*7,727*	**1**	**0.675**	*7,045*
Age Head of HH (pilot data)	**3**	0.593	*437*	**2**	0.625	*391*	**2**	**0.665**	*366*
Race/ethnicity	**2**	**0.619**	*8,880*	**3**	0.608	*6,509*	**3**	0.608	*5,894*
Marital status	**4**	0.467	*10,480*	**4**	0.502	*7,727*	**4**	**0.546**	*7,045*
Household income	**5**	0.445	*11,162*	**5**	0.456	*8,234*	**5**	**0.470**	*7,484*
Children in household	**7**	0.357	*11,537*	**7**	0.367	*8,496*	**6**	**0.386**	*7,716*
Education of Head of HH	**6**	0.365	*9,302*	**6**	0.379	*6,839*	**7**	**0.385**	*6,198*
Number of adults	**8**	0.261	*10,480*	**8**	0.281	*7,727*	**8**	**0.302**	*7,045*

* All correlations are significant at p<.0001

Figures 2 through 4 show the accuracy rates for selected ancillary data. Accuracy is defined as the ability of ancillary data to predict the self-report survey data. As a guide for how to interpret the charts, consider first the example of "all" panel recruits for the ancillary variable of home ownership, as shown in Figure 2. Among the adults for whom the ancillary data predicted home ownership (instead of renting), the ancillary data correctly predicted home ownership in 93.3% of the recruited adults. Only 6.7% of the predicted "home owners" turned out to be renters (per the survey data). The ancillary data, therefore, could provide an efficient method for sample targeting home owners, which is consistent with the use of infoUSA and like firms that serve as data warehouses for credit-rating purposes.

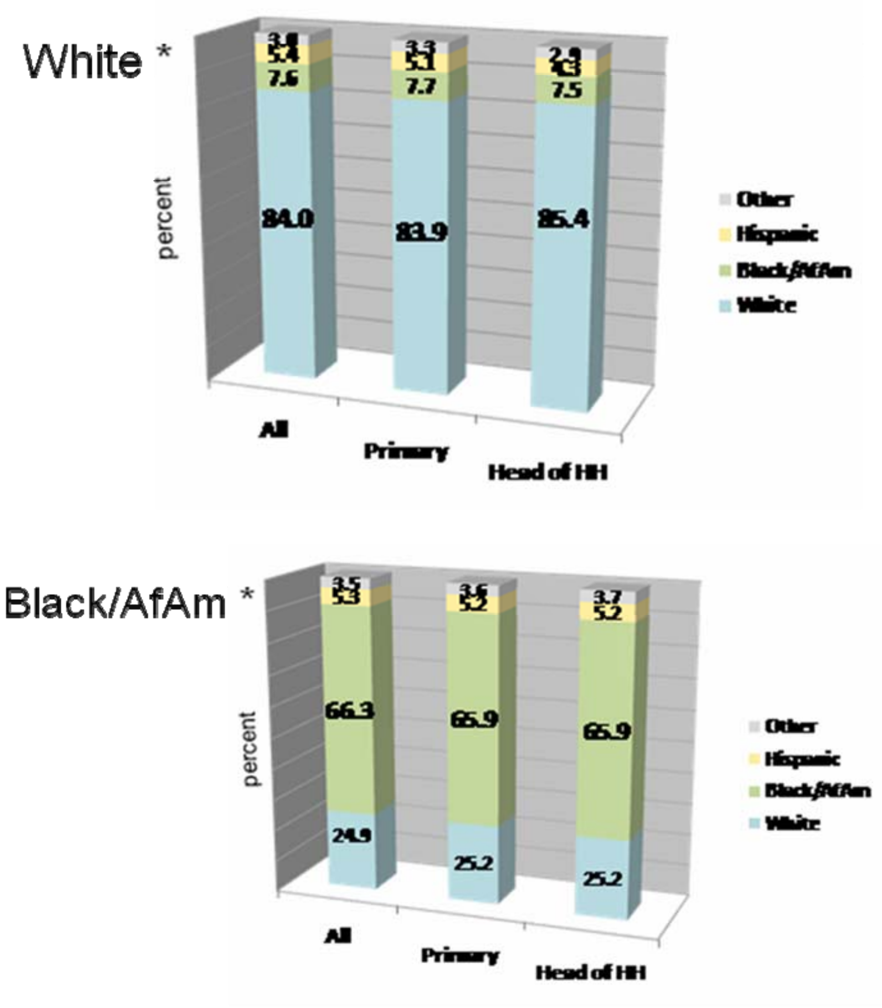
Figure 2: Accuracy Rates for Home

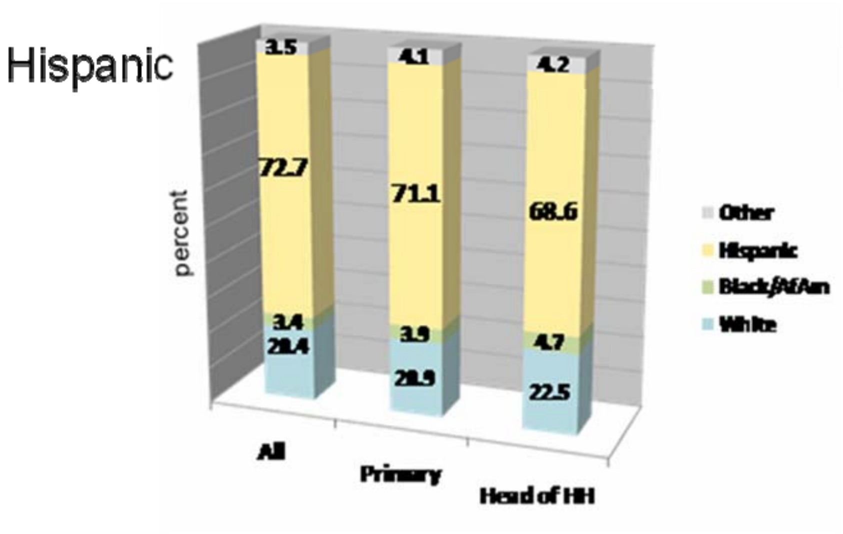


Ownership

Figure 3 displays the accuracy rates for race/ethnicity. The ancillary data accurately predicted the racial and ethnic identification of recruited panelists in approximately two out of three cases for African Americans and almost three of four recruited Hispanics.

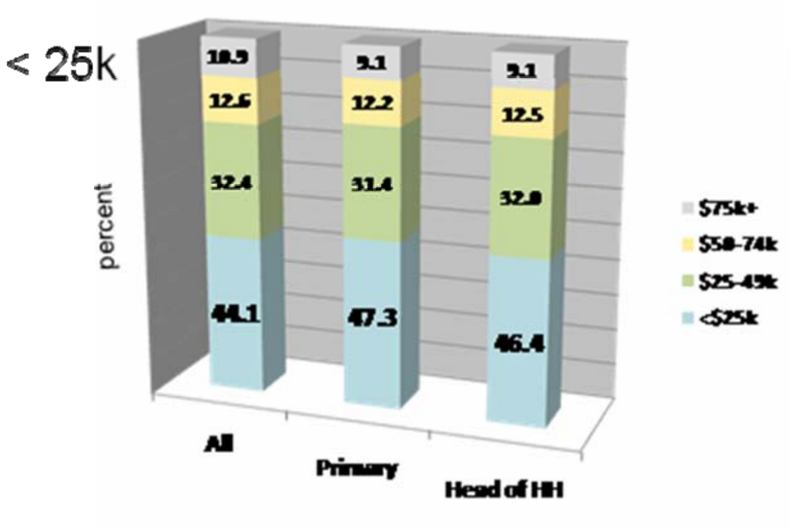
Figure 3: Accuracy Rates by Race/Ethnicity

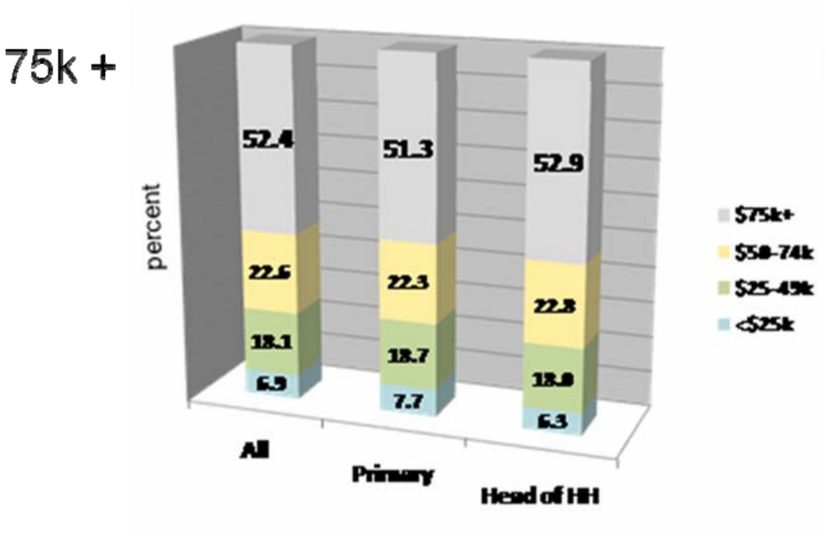




The ancillary data were less accurate in predicting household income. For the lower-income households (less than $25,000 per year), the accuracy rate was less than 50%. For the higher-income households, the accuracy rate was slightly higher (at 52%), as shown in Figure 4.

Figure 4: Accuracy Rates for Household Income: Less than $25,000 and $75,000 and Higher per Year





5. Analysis of Ancillary Data for Non-Response Tests

For a subset of approximately 4,000 KN recruited households, all recruited during the Spring 2009 ABS recruitment, we used the ancillary data to compare all the invited sample units to the 4,000 recruited households. For this analysis, only ancillary data are used – no self-reported survey data. We assume that error in the ancillary data is randomly distributed across the “invited” sample and subset of invited sample units that make up the “recruited” sample.

The non-response bias test consists of comparing the sample composition of the invited sample to the subset of sample units that agreed to join KnowledgePanel. Evidence of self-selection bias would be clear if, for instance, the frequency distribution of the invited sample is substantially different from that of the recruited sample for demographic characteristics such as educational obtainment, race/ethnicity, and household income. The results of that comparison for these three characteristics are displayed in Figures 5 through 7.

We found substantial alignment of the recruited households compared to the invited sample.

Figure 5: Educational Obtainment: Frequency Distribution of the Total Invited Sample to the Subset of Recruited KnowledgePanel Adults

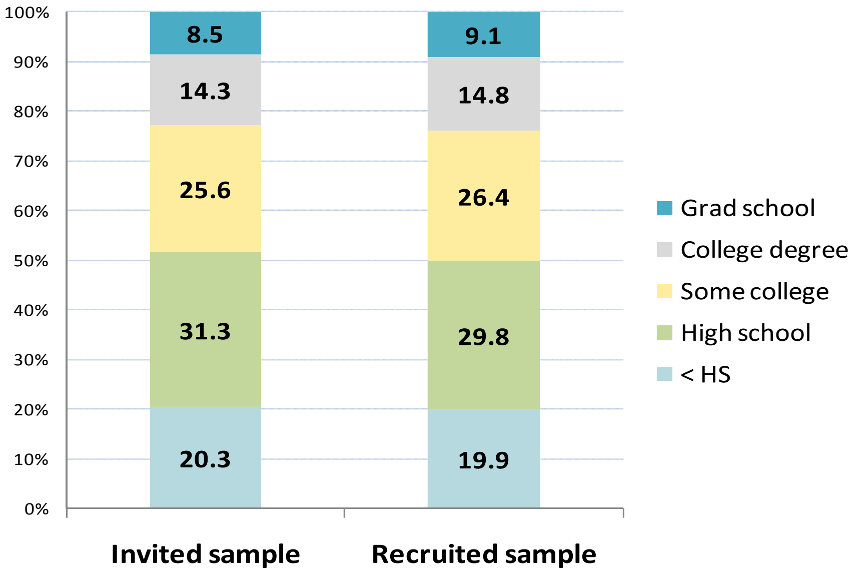


Figure 6: Race/Ethnicity: Frequency Distribution of the Total Invited Sample to the Subset of Recruited KnowledgePanel Adults

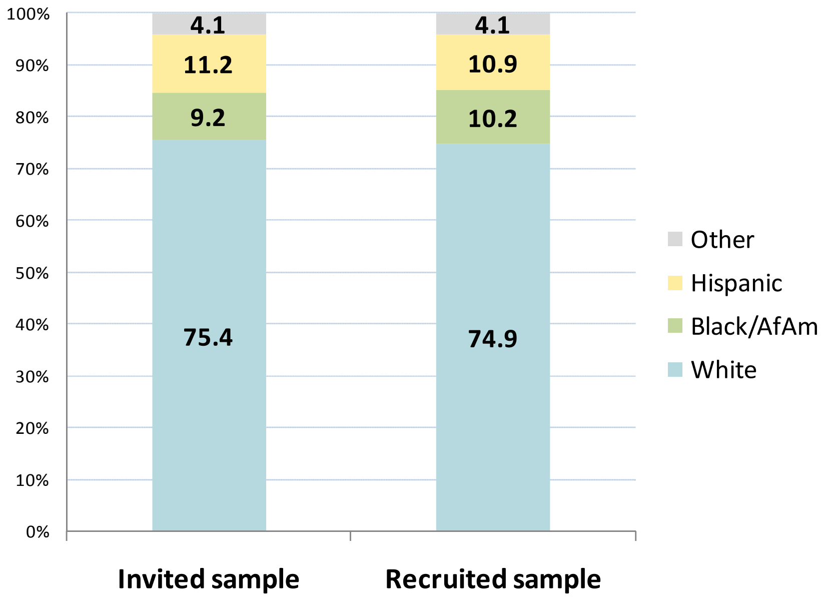
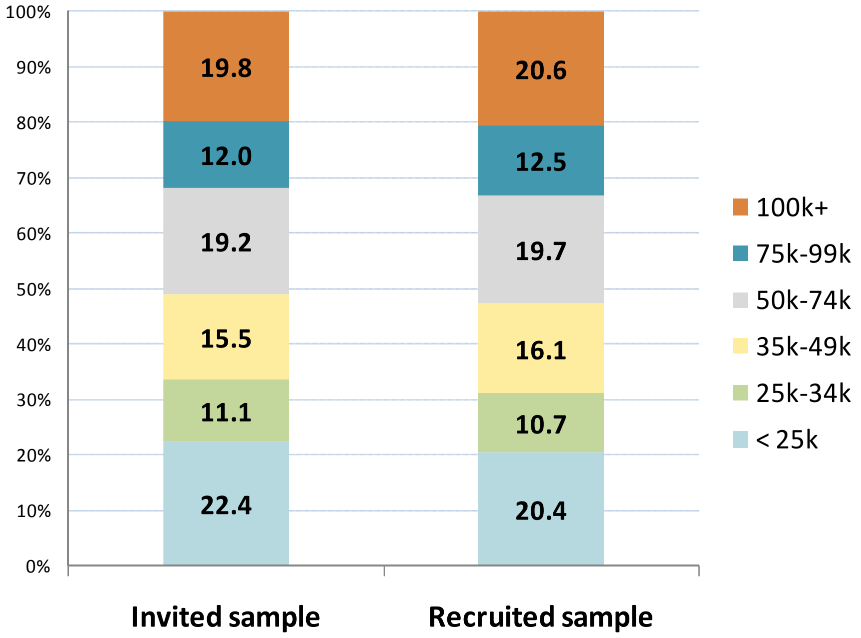


Figure 7: Household Income: Frequency Distribution of the Total Invited Sample to the Subset of Recruited KnowledgePanel Adults



6. Conclusions

1. Ancillary information attached to ABS samples has value for examining non-response in surveys and for improving the efficiency of complex sample stratification designs. Self-selection biases were not evident in the KN panel sample based on the ABS frame.

2. Analysis using additional variables holds promise for future non-response measurement and sample stratification.

3. Ancillary information appears to correlate best with head-of-household information.

4. Ancillary information may be useful for mail strategies targeting homeowners, race/ethnic groups, household income extremes, and perhaps with age groups and other groups (pending further research).

Acknowledgements

We acknowledge the active data analysis support and insights of Erlina Hendarwan of Knowledge Networks.

References

Dennis, J.M. 2010a. KnowledgePanel Processes and Procedures Contributing to Sample Representativeness and Tests for Self-Selection Bias. Available at www.knowledgenetworks.com/ganp/reviewer-info.com.

Dennis, J.M. 2010b. Summary of KnowledgePanel® Design. Available at www.knowledgenetworks.com/ganp/reviewer-info.com.

DiSogra, C. 2010. Maximizing a Stratified ABS Sampling Frame for Nationwide Mail Recruitment of a Probability-Based Online Panel. Available at www.knowledgenetworks.com/ganp/reviewer-info.com.