Assessing Coverage and Accuracy of Population Subgroups using an Address-Based Sample Frame

Kelly Dixon[1], Mike Kwanisai[1], Dan Estersohn[1], Al Tupek[1],
Linda Piekarski[2], Missy Mosher[2], Jessica Smith[2]

[1]Arbitron Inc, 9705 Patuxent Woods Dr., Columbia, MD 21046
[2]Survey Sampling International, 6 Research Dr., Shelton, CT 06484

Abstract
Arbitron obtains hundreds of thousands of records annually from Survey Sampling International (SSI) for purposes of selecting household samples. The primary information that Arbitron gets from SSI is either a phone number or an address (sample point) and some geographic descriptive information such as county or subcounty. SSI has the ability to provide additional information about the sample points; including such things as name, age and race/ethnicity of the householder, existence of certain persons of certain age group or gender in the household. Achieving a sample of respondents that satisfies demographic and geographic proportionality is one of the main goals for Arbitron, since radio listening does vary by these characteristics. We analyzed the usefulness of the SSI auxiliary information to find Hispanic and young households in order to sample them at a rate that yields a proportional sample. We compared the demographic information from thousands of Arbitron respondents to the SSI frame information. We report the proportion and types of matches and non-matches by household characteristic. We discuss data quality metrics in terms of accuracy and coverage and the importance of each to achieve our survey's goals.

Key Words: stratified sampling, coverage, accuracy, misclassification error

1. Introduction

Arbitron (NYSE: ARB) obtains hundreds of thousands of records annually from Survey Sampling International (SSI) for purposes of selecting samples of households for our diary and Portable People Meter (PPM) service. The primary information that Arbitron gets from SSI is a phone number or an address (sample point) and geographic descriptive information such as county or subcounty. SSI can however provide additional information about the addresses or sample points which Arbitron may, if accurate, potentially take advantage of in its sampling procedures. This sample point information includes such things as name, age and race/ethnicity of the householder, age and gender of persons in household, presence and number of children in household, presence of young adults, household size and income. The SSI household data[1] includes flags to identify if the householder surname is on a Hispanic surname list.

[1] We understand that SSI does not own the data but obtains the data through its vendors. For the purpose of this report we will refer to the data as SSI's meaning the data that SSI could provide though its vendors.
[2] Arbitron media markets define a collection of counties within major metropolitan areas that are

The household-level information is obtained from linking and matching data from various public records and marketing lists. As a result, the quality of SSI's data depends on how accurate, complete and current the information from the various sources is, and also how the information was linked and how non-matching cases were resolved. Although SSI's household data may not be perfect, it is worthwhile to assess its quality and fitness for Arbitron's use for sampling purposes. Scarborough, another media research provider, has been successfully using SSI's information on Hispanics to improve the proportionality of Hispanics in its sample.

Arbitron has traditionally used a proportional sample design to measure radio audiences. Since radio listening tends to vary by both demographic and geographic characteristics, Arbitron has always strived to select samples that are representative across these characteristics. Achieving better demographic representation depends on several factors such as market [2] and demographic characteristics, response rates, sample size, sampling procedure, targeted incentives and treatment among other factors.

Arbitron has two types of services (PPM and Diary) for producing radio ratings to its clients. The PPM (Portable People Meter) service is used in 48 of the largest metros, while the Diary service is used in the remaining metros throughout the country. The PPM service consists of a panel of survey participants who are sampled from a geographically stratified address based frame and agree to wear the PPM for a period of up to two years. Initial contact to the household is made by phone and/or mail and repeated contact and follow-up attempts are made. The (PPM) is a small apparatus, the size of a pager, which transmits a cellular signal containing the listening and motion data each night to Arbitron. The listening data is tabulated into the ratings based on whether the panelist actually wore their meter long enough during the course of the day. The Diary service samples households from a dual frame (RDD and address-based frames) that are also stratified geographically. Households are recruited via phone and mail attempts. Households that agree to participate fill out paper diaries of their radio listening for a period of seven days. The survey period for the Diary service is twelve weeks, and each market is measured with an independent sample either two or four times a year.

Both services face challenges in producing samples that are demographically representative of the population due to the national trend of lower response among survey participants in general; and amongst younger and more ethnic populations specifically. Arbitron offers numerous incentives to garner participation to younger and ethnic households, but still falls short of its proportionality goals.

Currently Arbitron uses a screener during recruitment to gather important information about the households after they have been selected for sampling purposes. Gathering initial household characteristic information during recruitment is a process that is time-consuming and expensive and may lower the overall response rate. It is therefore worthwhile to assess the quality of SSI's data to investigate and explore better sampling frames and efficient sampling procedures to adopt.

[2] Arbitron media markets define a collection of counties within major metropolitan areas that are used to categorize radio ratings. These closely resemble the more well-known designated market areas, or DMA's, that are similarly defined by the Nielsen Company to categorize television ratings.

The objective of this study was to investigate the quality of household characteristics that SSI can append to sampled addresses. Finding the information dependable for identifying certain race/ethnicity and age groups encourages development of sampling strategies that yield a more proportional sample. Specifically, we wanted to know if SSI data can provide reliable information on Hispanic and young adult (persons age 18-34) households. With reliable information better sampling methods for targeting households with these characteristics in the population could be chosen prior to recruitment. Also, the early identification of households especially those that are difficult to recruit will help better target material requirements and incentives in future surveys.

2. Methodology

Our study evaluated data that SSI could provide on 24,698 addresses from 10 PPM and 8 diary markets. We purposefully selected these markets because they include high proportions of Hispanics and college students. Such demos tend to provide the greatest proportionality challenges and thus obtaining better quality additional household data especially for such demos could suggest more efficient sampling methodologies to adopt.

The addresses included households with known characteristics due to their participation in Arbitron surveys. To test accuracy of SSI's information on these addresses, SSI was supplied with a list of the addresses and tasked to both link (match) the addresses and append household information using their databases. Table 1 shows the distribution of linked and unlinked addresses. Appendix A shows the distribution of linked and unlinked addresses by market.

Table 1. Distribution of linked and unlinked addresses

			Diary	PPM	Total
Total			13,978	10,720	24,698
Unlinked			985	1,265	2,250
Linked	Returned Diaries	Hispanic market	3,223	9,455	22,448
		Non-Hispanic market	5,750		
	Total Returned Diaries		8,973		
	Unreturned Diaries		4,020		
	Total Linked Diaries		12,993		

2.1 Measuring Quality

The demographic and household characteristics information from current PPM panelists and diary keepers as reported to Arbitron was compared to SSI appended data by race/ethnicity and age groups. We used coverage, accuracy, total agreement and misclassification error rates to assess data quality.

Table 2 shows the way in which comparisons were made between the two data sets on each household characteristic of interest. The SSI data included households with linked addresses but unknown household characteristic. The cell entries in Table 2 cell entries represent number of households as defined by Arbitron and SSI.

Table 2: Comparing SSI and Arbitron household data

		SSI (Test Data)			
		Unknown HH Data	No	Yes	Total
Arbitron (True Data)	No	a	b	c	a+b+c
	Yes	d	e	f	d+e+f
	Total	a+d	b+e	c+f	a+b+c+d+e+f

2.2 Terminology and definitions

Accuracy was defined as the percentage of households that SSI indicated has a specific characteristic and were also reported as such according to Arbitron data. Of the addresses that SSI indicated as Hispanic households, what percentage is Hispanic according to Arbitron reported data? This was computed as $\frac{f}{c+f}$.

Coverage was defined as the percentage of households that Arbitron reported to have a certain characteristic that are also indicated as such according to SSI data. This is a measure of sensitivity. This was computed as $\frac{f}{d+e+f}$.

Overall agreement was defined as the percentage of households classified as with or without a specific characteristic by both data sources. This was computed as $\frac{a+b+f}{a+b+c+d+e+f}$.

We focused on two misclassification errors; a false negative and a misclassification from linked addresses with unknown household information, which in practice would result in falsely missing these households with the characteristic of interest. We distinguish the two by referring to them and "problem negative" and "problem unknown".

Problem negative is the percentage of households SSI indicates as not having a characteristic of interest when they actually have according to reported Arbitron data. We calculated this as $\frac{e}{b+e}$.

Problem unknown is the percentage of households SSI does not have household data to indicate they had the characteristic of interest when they actually have according to reported Arbitron data. We calculated this as $\frac{d}{a+d}$.

An alternative measure of quality of binary classification is Matthews correlation coefficient (MCC) which takes into account true and false positives and negatives. MCC can be expressed as

$$MCC = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

where a = true positives, b=false positives, c=false negatives and d=true negatives.

3. Results

In this section we present the results of the SSI appended data using four measures defined in the previous section. Our assessment mostly focused on accuracy and coverage of the SSI data with respect to households with Hispanics, 18 to 34 year olds and 55+ year olds. We also present results of comparisons made on unlinked and linked addresses to test if certain households are linked at higher or lower rates by SSI.

SSI linked 91% of the addresses that were provided. Appendix A shows the percentage of linked addresses by market. SSI was able to append household characteristic data on 92% of the linked addresses.

3.1 Hispanic Households

Table 3 shows accuracy, coverage, agreement and misclassification rates for Hispanic households. A high accuracy rate of 88% and a coverage rate of 64% for Hispanic households were obtained. Overall agreement rate was 88% and very high.

Table 3: Accuracy, coverage and misclassification rates for Hispanics households

	PPM	Diary	All
Accuracy	91%	80%	88%
Coverage	63%	69%	64%
Agreement	88%	89%	88%
Problem negatives	10%	7%	9%
Problem unknown	32%	22%	29%

The high accuracy rates imply that the SSI data accurately Hispanic households 88% percent of the time. The coverage rate was 64% and this means that while SSI data could help targeting Hispanic households with 88% accuracy, it only covers 64% of the Hispanic household population. 29% of households with unknown characteristics were Hispanic while only 9% that were falsely flagged by SSI as not Hispanic households.

Appendix B shows accuracy, coverage and misclassification rates of households with Hispanics by market. The market-level results show that although the rates are generally comparable, a few markets have lower accuracy and coverage rates.

3.2 Households with Persons 18-34 year olds

Table 4 shows accuracy, coverage, agreement and misclassification rates for households with persons aged 18 to 34 years.. For diary markets, the results are from linked returned diaries only. Unlike results for Hispanic households the results indicate less reliable results pertaining households with young adults. Although these results are not

encouraging, the SSI appended information could nevertheless be useful, maybe in conjunction with other frames to help target young adults.

Table 4: Accuracy, coverage and misclassification rates for households with 18-34 year olds

	PPM	Diary	All
Accuracy	69%	51%	61%
Coverage	47%	47%	47%
Agreement	70%	74%	72%
Problem negatives	28%	17%	22%
Problem unknown	43%	37%	40%

Appendix C shows accuracy, coverage and misclassification rates of households with young adults by market. The market-level results show about the same rates across markets with a few markets showing lower results.

3.3 Households with Persons 55+ year olds

Table 5 shows accuracy, coverage, agreement and misclassification rates for households with persons aged 55 years and older. The accuracy and coverage rages are high and comparable to those of Hispanic households. Thus SSI data seem to be relatives accurate and has better coverage for Hispanic and 55+ households.

Table 5: Accuracy, coverage and misclassification rates for households with 55+ year olds Households with 55+ year olds

	PPM	Diary	All
Accuracy	86%	89%	88%
Coverage	74%	80%	77%
Agreement	82%	83%	82%
Problem negatives	18%	20%	19%
Problem unknown	38%	47%	42%

Appendix D shows accuracy, coverage and misclassification rates of households with persons 55 years and older by market. The market-level results show relatively similarly good accuracy and coverage rates across all markets.

3.4 Households with other demographics

Table 6 below shows rates for other demographics. The SSI data was not as good with these demographics with households with Blacks having the lowest coverage of 4%. Accuracy rates were sometimes high the low coverage and high misclassification rates were not good.

Table 6: Households with other demographics

	Accuracy	Coverage	Agreement	Problem negatives	Problem unknown
Children 6-17	68%	34%	78%	20%	23%
Male 6+	90%	65%	66%	69%	75%
Female 6+	96%	64%	66%	74%	77%
Male 18-34	59%	35%	77%	19%	29%
Female 18-34	61%	39%	77%	19%	28%
Black	73%	4%	83%	16%	22%
Asian	63%	29%	97%	3%	3%
Other Race	79%	80%	74%	31%	53%
Adults 18-24	61%	34%	82%	15%	17%

3.5 Linked and Unlinked Addresses

Not all addresses were linked by SSI. Addresses were unlinked when SSI could not identify an Arbitron supplied address on their database with confidence. Given that SSI did not link all addresses, we wanted to assess if there was any bias in terms of addresses that SSI could link. We limited our analyses to focus on bias with respect to Hispanic, 18 to 34, and 55+ year old households. Table 7 summarizes test results of linked against unlinked addresses.

Table 7: Percentage of addresses within known, unknown and unlinked households

	Market	Linked & Known	Linked & Unknown	Unlinked
Hispanic	PPM	27%	32%	33%
	Diary	22%	22%	37%
	All	26%	29%	34%
18-34 years	PPM	40%	43%	42%
	Diary	26%	37%	36%
	All	33%	40%	40%
55+ years	PPM	48%	38%	39%
	Diary	59%	47%	48%
	All	53%	42%	41%

Except to Hispanic households in diary markets, we did not find any significant difference between percentages in unlinked and unknown addresses. Comparing percentages in known and unknown addresses, we found that unknown households had fewer 55+ year olds and more 18-34 year olds compared to addresses the SSI had known households characteristic.

4. Conclusions

We learned that SSI data could be effective in identifying households by characteristics, with the effectiveness varying by characteristic.

Hispanic households

We learned that the identifying of Hispanic households using surname is effective. SSI used Hispanic surname indicators to flag Hispanic households and their results agreed to a great extent with what we obtained from Hispanic respondents. This conclusion is in line with what we expected about Hispanic surname as a reliable household identifier. From households flagged by SSI as Hispanic, 88% of them were correctly identified as Hispanic when compared to Arbitron respondent data.

Households with Persons 18 to 34

We learned that SSI data was less efficient in identifying households by presence of 18 to 34 persons compared to when identifying Hispanic households. The SSI data correctly identified 61% of households that it considered as having an 18 to 34 year old. We note that our data mostly covered college and Hispanic markets so results could be better in markets without these characteristics.

Households with Persons 55+

Although it was not a specific objective of this study, we found that the accuracy, coverage and agreement rates for addresses with 55+ year olds are high. The accuracy rates were better than the accuracy rates for households with Persons 18 to 34. This group is typically over-represented in Arbitron samples because they are better survey respondents than younger households and are better covered by our sample frames. This information can also help us devise a sampling strategy to select fewer of these housheolds

Misclassification Error

We learned that misclassification errors were generally small but do exist. These also vary by characteristic (i.e. more misclassification errors for young households than older ones.) Any sampling strategy would need to account for classification errors

Missing Household information (Unknown)

We found that addresses with missing frame information (or unknown households) have a higher proportion of younger adults. Oversampling these addresses could target 18-34 year olds. We found that even the absence of frame characteristics yields information that can be used for sampling.

5. Recommendations

Arbitron is encouraged by the findings of this study and plans to pursue testing and simulations of sample designs using frame information from SSI for further stratification of its sample frames. Since SSI data seems reliable especially in identifying Hispanic households, we should devise a methodology to utilize this information in our sampling procedures prior to first contacting the respondents. Using SSI data together with block group, tract or county-level US Census, ACS or other data sources may enhance accuracy and sample design.

We envision that the frame information can have utility aside from sample design. SSI data could be used to enhance or replace collecting initial demographic data for sample planning to improve survey process efficiency. We can also investigate using SSI data together with different treatments to target demographics with low response rates.
Further analyses of SSI data on other markets and households characteristics should be conducted to ascertain the full strength of the data by markets and household characteristic. Finally, listening analyses by linked and unlinked addresses will be beneficial to assure clients that potentially over-sampling linked addresses will not affect ratings. Our hypothesis is that there will be little to no differences between the two groups.

6. Limitations

There are several limitations and assumptions made in the study that should be noted when considering the results:

Unlinked addresses. There were some addresses that Arbitron had in its set of respondents that SSI could not match (using an exact match) to their database. This could have been due to the address truly not existing on SSI's frame or an error during the matching. The household characteristics of these addresses may potentially be different from that of persons from linked addresses. We assumed results from the linked data mirrors what we would obtain from the unlinked data. Thus we assumed systematic patterns or distributions between the linked and unlinked data on the characteristics of interest. We address this with additional analyses in the report. For the test data, a comparison analysis on linked against unlinked data on households with Hispanics and 18 to 34 year olds did not show significant differences on these two characteristics.

Respondents are correct. The Arbitron data was used as the "gold standard" thus we assumed that the respondents' answers reflect the truth about the household.

Returned diaries and installed panelists. We only compared frame information from persons who were installed PPM panelists at one point between January 2011 and June 2011, or persons who returned diaries. We could also look at non-responders on various characteristics. To the extent that the likelihood to match to the SSI lists is different, our results could vary. This is likely to be a larger problem for the diary service than in the PPM service. In PPM, all panelists must agree to participate before a household is installed. Therefore, the sample of panelists within a household is likely to be complete. In the Diary service, while we mail diaries to everyone, not everyone returns a diary. Therefore, there may be situations where SSI indicates the presence of persons 18-34, our records show that no one at that address returned a diary.

Time lag. If we decide to stratify on SSI information, there will be a time lag between sample selection and returned diaries or installed PPM panelists. In this study, the time lag was reversed. The respondent household data that we used was generally collected several months prior to matching to SSI. So instead of the SSI information lagging the respondent information, the study has respondent information lagging the SSI information. We are unsure how this may impact the results. but it is likely that the match rates will overstate what we are able to achieve in a production environment. A test that would come close to simulating the actual time lag is planned for spring 2012.

Unknown versus no information. SSI used flags to identify existence and non-existence of persons of specific age groups or race/ethnicity. A default flag was used for non-existence of certain individuals in a household. SSI did not differentiate if the non-existence flag meant that information was unknown or that information was known but non-existent. For example the ADULT_AGE variable used "0" both as a default for unknown age and also to represent a household without a person in the age group.

Purposeful selection of markets. This was not a random selection of markets. We purposefully selected markets that had proportionality issues with demos of interest such as minorities, college students and young adults. Because of this, we think that results for the remainder of markets could be different.

References

Carugo, O. (2007). Detailed estimation of bioinformatics prediction reliability through the Fragmented Prediction Performance Plots, *BMC Bioinformatics*, **8**:380.

Khare, M., Ezzati-Rice T.M., Battaglia, M.P. and Zell, E.R. (2001). An assessment of misclassification error in provider-reported vaccination histories. *Proceedings of the Annual Meeting of the American Statistical Association*, August 5-9.

Appendix A. Distribution of linked and unlinked addresses by market

Market Name	Market Code	Market Type	Hispanic Market	Linked	Not linked	Total	% Linked
Fresno	089	Diary	Yes	2,247	215	2,462	91%
Amarillo, TX	147	Diary	Yes	1,031	81	1,112	93%
Boise	229	Diary	Yes	1,016	73	1,089	93%
Danbury, CT	593	Diary	Yes	518	39	557	93%
Omaha	085	Diary		2,255	117	2,372	95%
Birmingham	095	Diary		2,832	246	3,078	92%
Lansing	195	Diary		1,823	105	1,928	95%
Gainesville	550	Diary		1,271	109	1,380	92%
All Diary markets				12,993	985	13,978	93%
Los Angeles	003	PPM	Yes	1,677	219	1,896	88%
Chicago	005	PPM	Yes	1,216	286	1,502	81%
Washington, DC	015	PPM	Yes	940	88	1,028	91%
Dallas-Ft. Worth	024	PPM	Yes	905	94	999	91%
Denver	035	PPM	Yes	853	96	949	90%
Sacramento	065	PPM	Yes	758	65	823	92%
Raleigh	115	PPM	Yes	482	47	529	91%
Miami	429	PPM	Yes	1,165	215	1,380	84%
Pittsburgh, PA	023	PPM		911	101	1,012	90%
Columbus, OH	045	PPM		548	54	602	91%
All PPM markets				9,455	1,265	10,720	88%
All markets				22,448	2,250	24,698	91%

Appendix B. Accuracy, coverage and misclassification rates of households with Hispanics by market (Hispanic markets only)

Market Name	Market Code	Accuracy	Coverage	Agreement	Problem negatives	Problem unknown
Los Angeles	003	90%	65%	84%	14%	47%
Chicago	005	91%	64%	93%	5%	14%
Washington, DC	015	82%	46%	90%	9%	11%
Dallas-Ft. Worth	024	93%	67%	90%	8%	29%
Denver	035	90%	56%	88%	10%	32%
Sacramento	065	83%	52%	88%	10%	24%
Fresno	089	87%	74%	85%	12%	38%
Raleigh	115	92%	62%	95%	3%	18%
Amarillo, TX	147	52%	59%	89%	4%	13%
Boise	229	59%	42%	93%	4%	9%
Miami	429	96%	68%	83%	20%	53%
Danbury, CT	593	62%	46%	94%	4%	4%
All Markets		88%	64%	88%	9%	29%

Appendix C. Accuracy, coverage and misclassification rates of households with 18 to 34 year olds by market

Market Name	Market Code	Accuracy	Coverage	Agreement	Problem negatives	Problem unknown
Los Angeles	003	74%	37%	67%	34%	45%
Chicago	005	68%	49%	71%	27%	39%
Washington, DC	015	71%	45%	71%	28%	36%
Pittsburgh, PA	023	66%	50%	72%	25%	40%
Dallas-Ft. Worth	024	69%	48%	71%	27%	41%
Denver	035	65%	50%	70%	24%	60%
Columbus, OH	045	66%	65%	74%	19%	44%
Sacramento	065	71%	41%	68%	31%	46%
Omaha	085	52%	50%	75%	16%	36%
Fresno	089	59%	35%	67%	29%	43%
Birmingham	095	49%	48%	75%	15%	32%
Raleigh	115	64%	60%	74%	19%	45%
Amarillo, TX	147	48%	48%	74%	15%	33%
Lansing	195	47%	59%	75%	11%	38%
Boise	229	46%	42%	70%	18%	47%
Miami	429	72%	47%	70%	30%	41%
Gainesville	550	48%	43%	79%	12%	32%
Danbury, CT	593	59%	62%	80%	12%	17%
All Markets		61%	47%	72%	22%	40%

Appendix D. Accuracy, coverage and misclassification rates of households with 55+ year olds by market

Market Name	Market Code	Accuracy	Coverage	Agreement	Problem negatives	Problem unknown
Los Angeles	003	85%	70%	81%	19%	35%
Chicago	005	83%	76%	81%	16%	43%
Washington, DC	015	86%	68%	80%	18%	49%
Pittsburgh, PA	023	88%	81%	83%	18%	38%
Dallas-Ft. Worth	024	83%	75%	84%	13%	29%
Denver	035	88%	80%	86%	13%	37%
Columbus, OH	045	87%	81%	86%	13%	37%
Sacramento	065	85%	76%	82%	18%	32%
Omaha	085	90%	84%	86%	14%	50%
Fresno	089	86%	75%	81%	19%	38%
Birmingham	095	90%	82%	83%	22%	50%
Raleigh	115	85%	75%	85%	12%	42%
Amarillo, TX	147	92%	80%	83%	21%	48%
Lansing	195	89%	82%	82%	23%	47%
Boise	229	87%	76%	80%	22%	43%
Miami	429	87%	68%	77%	29%	38%
Gainesville	550	91%	79%	82%	24%	54%
Danbury, CT	593	84%	80%	83%	16%	43%
All Markets		88%	77%	82%	19%	42%