Modeling the Need for Traditional vs. Commercially-Available Address Listings for In-Person Surveys: Results from a National Validation of Addresses

Colm O'Muircheartaigh[1], Ned English[2], Michael Latterner[2]
Stephanie Eckman[2], Katie Dekker[2]
[1]Harris School of Public Policy Studies at the University of Chicago, 1155 E. 60th Street, Chicago, IL 60637
[2]NORC at the University of Chicago, 55 E. Monroe Street, Chicago, IL 60603

Abstract

The use of commercially-available address lists as a replacement for traditional listing has been a topic of interest in recent years among survey research organizations due to the potential of substantial cost savings. Of concern has been the possibility that the coverage of commercially-available listings is inferior when compared to the traditional "gold standard", at least in specific environments. NORC has employed the Valassis (formerly ADVO) commercial address list as the basis for area-probability studies in urban and suburban areas since 2003, while still conducting traditional listing in rural areas. The reason for the distinction has been the understanding that residents of rural areas tend to have PO BOX mail delivery, and thus do not have addresses that can be assigned directly to Census geographies through GIS (Geographic Information Systems) technology.

NORC conducted a field examination of Valassis address list on a nationally-representative sample of segments during the summer of 2008. The purpose of this address validation was to determine the coverage of the list generally, but with a specific interest in areas considered to be too rural for list sufficiency. Our paper first compares the results of our validation with what was expected using model that predicts list coverage based on a-priori information, such as urbanicity, population density, segment morphology, and other relevant attributes. Secondly, we present further refinements to our coverage prediction model, and demonstrate how it can be applied to any segment before listing begins. In so doing we argue that the suitability of a segment for commercial address lists or traditional listing contains a substantial random component. The third component of our research discusses the improvements that address list vendors have made in rural areas, and considers the implications for the ongoing need for traditional listing.

Key Words: Delivery Sequence File, Address-Based Sampling, General Social Survey

1. Introduction

NORC has been evaluating the status of field listing being the "gold standard" for household probability samples since 2001 (O'Muircheartaigh, Eckman, and Weiss, 2003). In so doing, NORC has been carrying out an examination of the

alternative approach of using the United States Postal Service delivery sequence file (DSF) as a basis for frame construction for area probability surveys (O'Muircheartaigh et al. 2006, O'Muircheartaigh et al. 2007). A key part of our research has been a national benchmark comparison, conducted between NORC and ISR at the University of Michigan, whose goal has been to provide a quantitative analysis describing the benefits and drawbacks of traditional listing vs. the USPS DSF for a national household sample.

In our first report on this research we compared a traditionally-listed housing unit (HU) frame to a USPS-based frame in the same set of areas (O'Muircheartaigh, Eckman, English, Lepkowski, and Heeringa, 2005). When discrepancies arose between the two frames, however, it was not possible to determine the source of the error, or which frame was more accurate.

Since then we have conducted additional field work in a subset of areas in order to reconcile the two frames with what was actually present in reality. Our field effort resulted in the creation of an edited "best" frame, which permitted the comparison of both the USPS and traditional listing-based approaches to a meaningful benchmark. Our second report (O'Muircheartaigh, English, Eckman, Upchurch, Garcia, and Lepkowski 2006) compared the "best" frame to the traditional and USPS-derived lists. We concluded that the USPS-derived list was a more effective representation of reality than the traditional list in most cases. One feature of our analysis was that *a priori* expectations as to which frame would be superior were frequently not correct.

We then took our analysis a step further by examining how well the "best" frame was represented by the T and U address frames in varying "real-world circumstances" (O'Muircheartaigh, English, and Eckman 2007). In so doing NORC developed a model that predicted coverage of traditional (T) and DSF-derived frames in a national sample of segments. Our model primarily considered Census data, such as median household income, race/ethnicity, and population density. It also included other derived variables, such as those describing the morphology of each segment, the presence of water features, and the percent of addresses in the associated ZIP code that were city style. Our research found that population density and the ratio of the USPS address count to the Census 2000 housing unit count predicted coverage of the USPS list. One shortcoming of our analysis was that it required known segments, which are often not present before frame construction.

It is clear that it would be valuable to survey researchers and practitioners to know which method, traditional vs. USPS, would be best for frame construction before starting a field project. As described in the literature there are considerable cost and coverage trade-offs associated with each method, and so it would be beneficial to identify the optimal approach. Clearly, one would anticipate using USPS delivery sequence file-derived lists in urban areas, and traditional listing in very rural areas, but most real-world situations are more complex.

Our current research had two primary goals. First, we develop a model to predict coverage using only variables known before fielding. In so doing our model predicts the overlap between the "best" or "B" frame representing reality with the USPS (U) and traditionally-listed (T) frames. Secondly, our research examines a selection of rural segments that previously required traditional listing in order to determine if the USPS databases have "caught up" in places that formerly did not contain city-style addresses. In so doing we compare database coverage in a variety of environments.

2. Methods

The *General Social Survey* (or "GSS") is a national area-probability in-person study conducted by NORC. For the current research we selected 34 of 446 GSS segments for in-person evaluation, which consisted of all 543 GSS segments minus 97 that did not contain any city-style addresses on the delivery sequence file due to their rural nature. We then acquired the USPS list from the Valassis vendor, known as the "ADVO" file, in the associated areas. After obtaining the ADVO file, we geocoded each address using the *MapMarker Plus* software package to determine their longitude and latitude. Following geocoding we were able to isolate the ADVO USPS list within the selected segments, which we define as the U frame.

Trained field staff then updated the USPS list within the selected segments during the summer of 2008. Staff were tasked with "confirming" or "rejecting" addresses, as well as adding addresses that were discovered in person not present on the list. We incorporated the input from the field staff to create the "best" or "B" frame. The B frame can be described as the U frame list plus any addresses that were added, minus any that were not present in the segment. The "B" frame is thus an enhanced and edited version of the U frame.

Following creation of the B frame, we utilized the RPart recursive partitioning algorithm and SEARCH binary segmentation program to help explain the intersections between B, T, U, and USPS (Sonquist, Baker, and Morgan, 1974). Both RPart and SEARCH function by dividing a data set though a series of binary splits into a mutually exclusive series of subgroups. The splits are selected at each step to maximize the variance explained by splitting the candidate set into two groups. We focus on the RPart results in the following results and discussion, as they were essentially the same as those from SEARCH.

The motivation of this research was to identify a model to permit *a priori* identification of types of areas that will have different frame quality. We chose the following variables that could be determined prior to the selection of specific target areas:

1. Population density, in persons per square mile

2. Land area, in square miles
3. Percent of housing units in urban areas, from Census 2000
4. Ratio of current U frame count to the Census 2000 unit count
5. Percent of housing units residing on blocks with city-style addresses, derived from Census 2000 Type of Enumeration Areas (TEA) code
6. Percent of housing units with city-style addresses on the ADVO file, calculated at the ZIP level
7. Percent of housing units occupied, from Census 2000
8. Median income, from Census 2000
9. Race and ethnicity measures, from Census 2000

In addition to validating the relationships between these variables and U list coverage, we also hoped to explain coverage in areas where the size of the U list has grown, remained constant, or declined in recent years. To capture any change, we considered the following explanatory variables: the ratio of the 2007 U frame count to the 2003 U frame count and the ratio of the 2003 U frame count to the 2000 Census housing unit count. We used the above set of variables to build two models in the RPart and SEARCH algorithms. Our first model was designed to predict "under-coverage" as measured by the percentage of units in the B frame that were also in the U frame, as denoted by the intersection B ∩ U. Our second model predicts "over-coverage" as measured by the percentage of units in the U frame not found in the B frame, as denoted by the intersection U not B.

3. Results

3.1 Selection of Segments

Table 1: Summary of segments selected

Category	*Description*	*Selected*	*Frame Size*
1	2003 U and 2007 U similar to Census 2000	11	363
2	2003 U higher than Census 2000, but 2007 U lower than Census 2000	1	5
3	2003 U lower than Census 2000, but 2007 U higher than Census 2000	4	8
4	2003 U higher than Census 2000, and 2007 U higher than Census 2000	5	41
5	2003 U lower than Census 2000, and 2007 U lower than Census 2000	4	9
6	Rural areas with limited U	5	13
7	Certainty selections	4	7
	Total	*34*	*446*

As mentioned in section 2, we drew a sample of 34 of 446 GSS segments, excluding those without any city-style addresses and thus no U list. We first selected 4 segments with certainty, as they had unusually high ratios of U to the census housing unit counts for both 2003 and 2007. We then stratified the urban segments using supplemental information such as change in the ratio of U to the Census housing unit counts between 2003 and 2007. Finally, we randomly selected a subset of rural segments containing some city-style USPS addresses. These segments are considered "rural" for this analysis because they have necessitated traditional listing for NORC surveys in the past due to under-coverage of the DSF file.

As shown in Table 1, the majority of segments had U counts similar to the census counts in both 2003 and 2007, as described as “generally stable” Category 1. Category 2 was designed to represent areas that likely experienced a decline in housing stock, a decrease in occupancy, or a decrease in coverage since 2003. Category 3 represents areas that have experienced the opposite e.g., an increase in stock, occupancy, or U coverage since 2003. Lastly, categories 4 and 5 represent areas with consistently high and low U coverage relative to the Census housing unit count respectively.

3.2 Intersection and Modeling Results

Table 2: Intersection Results by Category

Category	**Description**	**N**	**B ∩ U**	**U ∩ B**	**B ∩ USPS**	**U not B**
1	U03 ≈ Census, U07 ≈ Census	11	95.6%	97.9%	97.9%	2.1%
2	U03 > Census, U07 < Census	1	97.6%	99.8%	97.6%	0.2%
3	U03 < Census, U07 > Census	4	94.5%	98.0%	95.0%	2.0%
4	U03 > Census, U07 > Census	5	99.0%	98.3%	99.2%	1.7%
5	U03 < Census, U07 < Census	4	88.2%	97.6%	99.4%	2.4%
6	Rural areas	5	90.7%	94.1%	92.8%	5.9%
7	Certainty selections	4	80.8%	98.7%	91.2%	1.3%
Total		*34*	*95.0%*	*97.9%*	*97.5%*	*2.1%*

Table 2 presents weighted estimates of U and USPS coverage by selection category. Recall that U is defined as the DSF addresses that geocode inside a selected segment, while USPS would be all DSF addresses no matter their geocoded location. While the number of segments selected per category were relatively low, these estimates confirmed to basic expectations. First, B ∩ U for category 1 (stable) segments was very similar to the overall average. Second, areas that consistently displayed high ratios of U to the census housing unit count had the highest B ∩ U. Segments with consistently low ratios of U to the census

housing unit count (category 5) had relatively mediocre coverage. Finally, the rural segments had moderate to low B ∩ U.

What was not as expected was the relatively high coverage for category 2, which were segments that saw a drop U coverage since 2003. Two possible explanations for this result could be a lower occupancy rates in 2007 relative to 2003 or new home construction. Both units vacant for an extended time and new units yet to receive mail are not included on the delivery sequence file but would have been included in the B frame.

We were unsure as to how the next two categories would behave with respect to B ∩ U, those for which the U has increased in years past (category 3), and those where U consistently exceeded the census (category 4). In the areas for which the U count has increased over the last few years, we found moderate to high coverage with B ∩ U = 94.5% for category 3. Such a result would suggest that while the USPS frame may have been lagging in 2003 and has now effectively "caught up".

The results for our rural segments were as expected; on average, coverage was lower than that observed in suburban and urban segments. However, the USPS list performed surprisingly well in some of the rural segments. This raises the question of why some rural segments exhibited high coverage. To answer these and other questions as above, we utilized the RPart algorithm to distil the variables most responsible for the variation in U coverage.

Figure 2 presents the results of the recursive partitioning in a dendogram "tree diagram". The top node displays the overall B ∩ U estimate of 95% shown in table 2. The first split in the dendogram was on population density, indicating that density may be the most important variable in determining U coverage. Segments in tracts with less than 284 persons per square mile had on average inferior coverage than those with higher population density. Figure 2 uses the notation "U03" to describe the U count in 2003, "U07" to describe the U count in 2007, and "Census" to describe the Census 2000 housing unit count.

Of the 30 segments with a relatively high population density on the right side of the dendrogram, segments with a larger percentage of city style addresses were shown to have better U coverage. The percentage city style measure is derived from the USPS list as the share of housing units that were not PO Boxes or rural route boxes. This split follows the general assumption that segments with higher percentage city style addresses should have higher B ∩ U than those with a larger share of non city-style addresses.

Among the segments with a lower percentage city-style addresses, segments with the highest percentage city-style in that group (greater than 95%) had somewhat lower B ∩ U. This result is counter-intuitive, but it is based on a small sample

size (n = 2) and is derived from a ZIP code level measure which may not exactly fit a given segment.

Considering the right side of Figure 2, areas where the percent city-style addresses was greater than 95 were estimated to have 97% B ∩ U. We can see that segments which had a ratio of U03/Census housing unit count below 1 had a B ∩ U of 91% as compared to 97% for segments with a ratio of U03/Census housing unit count greater than 1. Among those with the U03/Census housing unit count below 1, those with a higher concentration African-American population were estimated to have higher coverage.

Among the 17 segments with a U03/Census greater than 1, the 15 segments that had a higher U07/U03 as derived from the 2003 and 2007 ADVO lists had a 98% B ∩ U. If one accepts U03/Census housing unit count and U07/U03 as measures of growth, it can be observed that segments with moderate or gradual growth may exhibit better U coverage than those with more rapid growth. One reason for this may be that the USPS list lags in areas with new construction and redevelopment. Figure 2 therefore details the variables that influence how well the U list captures the housing units found in the "B" frame reality.

Figure 3 presents a dendogram of factors influencing "over-coverage", or the percentage of U not in B. Over-coverage is undesirable in surveys because it implies loss of efficiency in both the sample and field operations. The overall percentage of addresses on the USPS list not verified in the field, our estimate of "over-coverage", was 2.1% as shown in the top node of figure 3. The dendogram in figure 3 first splits on the area of census tract. As with the first split calculated to explain B ∩ U, the split on area follows the urban vs. rural coverage findings from previous research, with large segments being analogous for rurality. Tracts that are less than 3.1 square miles in area were estimated to have less over-coverage than those larger than 3.1 square miles in area.

For the segments with larger areas, the subsequent split was made on percent African-American. Large tracts that were more than 2% African-American had on average more over-coverage than other segments.

For the segments with smaller areas, those with lower rates of city-style addresses (n = 2) tended to have more substantial over-coverage (B not U = 15%). This large figure was mainly driven by one of the segments, however, which was shown to have a U not B of approximately 24%.

Of the remaining 23 segments, those with an area greater than 0.42 miles were estimated to have the lowest rate of over-coverage. Specifically, only 0.6% of USPS addresses in segments with an area greater than 0.42 miles were not confirmed in the field, as compared to 3.2% for segments with smaller areas.

Figure 2: Dendrogram of B ∩ U

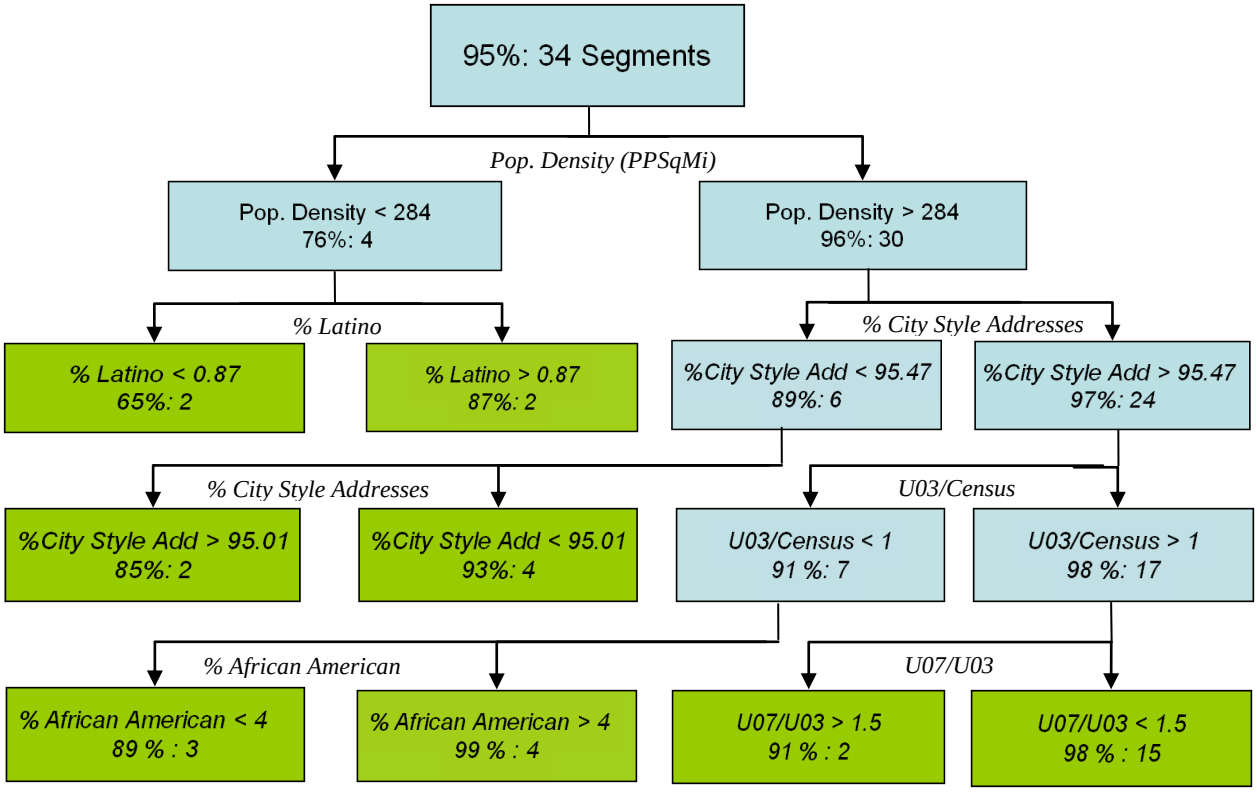
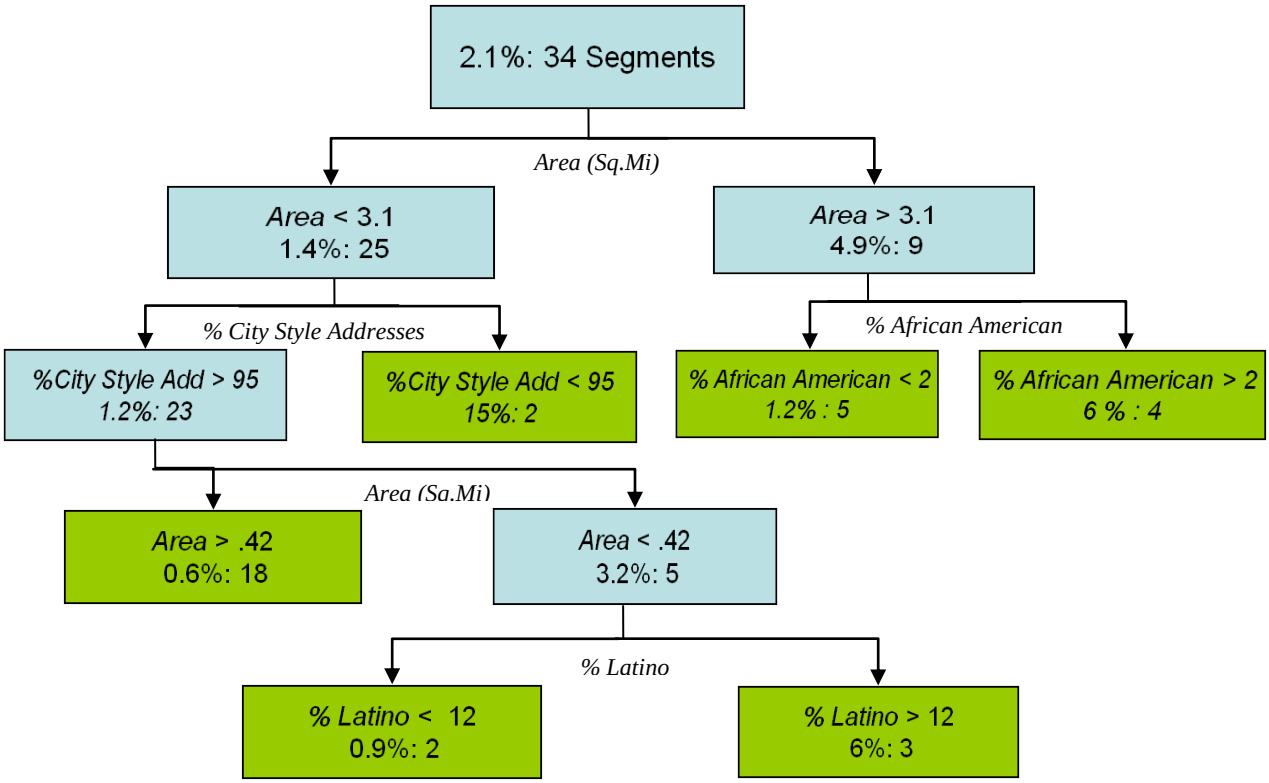


Figure 3: Dendrogram of U not B



4. Discussion and Conclusions

Our current evaluation demonstrated that the overall coverage of the U list was very good in segments with some DSF-based city-style addresses, with B ∩ U = 95%. We found that at the segment level, the overall coverage of the U list could be predicted by population density, the proportion of city-style addresses, in addition to information about U counts.

USPS over-coverage, as expressed by the percent U not B, was shown to be considerably more random than B ∩ U. While rurality in general does exert some influence on over-coverage, it is suspected that geocoding error and database update frequency are more important.

Our research does show that field listing may not be necessary in all "rural" segments, as the delivery-sequence file has improved over past performance in many areas. So, we feel these results illustrate that the USPS DSF has "caught up" in some areas through the conversion PO BOX addresses to city-style delivery. While the rural segments examined showed more over-coverage than others in the form of U not B, the percent B ∩ U would have been generally

acceptable for many studies. So, it may be worthwhile to ascertain the coverage of a delivery sequence-derived file in rural areas before resorting to a traditional listing.

Going forward we will be conducting a number of analyses to extend these results. We will first be running our models at the block-group level in order to resolve sub-segment differences in coverage. After that, we will focus our energy on learning more about the kinds of areas that tend to be missed by the T and U lists. Lastly, we will further investigate the degree to which rural areas are seeing city-style address conversion..

References

O'Muircheartaigh, C., Eckman, S., and Weiss, C. 2003. Traditional and enhanced field listing for probability sampling. *Proceedings of the Survey Research Methods Section, American Statistical Association.*.

O'Muircheartaigh, Colm, Stephanie Eckman, Ned English, James Lepkowski, and Steven Heeringa. *Comparison of Traditional Listings and USPS Address Database as a Frame for National Area Probability Samples.* Presented at American Association for Public Opinion Research Conference, May 2005, Miami Beach, Fl.

O'Muircheartaigh, C., English, N., Eckman, S., Upchurch, H., Garcia, E., and Lepkowski, J. 2006. Validating a sampling revolution: Benchmarking address lists against traditional listing. *Proceedings of the Survey Research Methods Section, American Statistical Association*.

O'Muircheartaigh, C., English, N., Eckman, S 2007. Predicting the relative quality of alternative sampling frames. *Proceedings of the Survey Research Methods Section, American Statistical Association*.

Sonquist, J. A., E. L. Baker, and J. N. Morgan. *Searching for Structure*. Revised edition. Ann Arbor: Institute for Social Research, The University of Michigan, 1974.