Modeling Coverage Error in Address Lists Due to Geocoding Error: The Impact on Survey Operations and Sampling

Lee Fiorio[1] and Jizhou Fu[1]

[1]NORC at the University of Chicago, 55 E. Monroe Street, Chicago, IL 60603

Abstract
Survey research organizations have been researching the use of extracts of the United States Postal Service delivery sequence file (DSF) as a replacement for traditional listing. Due to software limitations, individual housing units (HUs) on the DSF are sometimes errantly geocoded which can influence coverage properties of selected segments. NORC undertook a national listing effort in 2011 to augment the DSF in areas known to have limited coverage, such as rural areas and areas with new construction. We used an "enhanced" listing method, where the lister, using a handheld device, verifies and edits the DSF list geocoded to a designated segment. One benefit of enhanced listing is the ability to capture the geographic coordinates of each HU, thus providing data to further explore the nature of DSF coverage. We focus on a selection of rural and urban segments from the national listing effort. For addresses on the DSF but not found by the lister, we use logistic regression to model the likelihood of address-level geocoding error using DSF flags and census data from 2010. We also build an autologistic model (Besag, 1972) to account for spatially dependent data by incorporating spatial autocorrelation. Results indicate geocoding error occurrences are spatially dependent, and the probability of geocoding error is related to address characteristics such as drop delivery and address type as well as rural block characteristics including geographic area. Our model also demonstrates that low block-level DSF coverage is associated with geocoding error. Understanding the correlates of geocoding error in the DSF will increase listing efficiency and frame quality by allowing the identification of areas with the most limited DSF coverage that will require listing for sampling frame construction.

Key Words: Address-based samples, geocoding error, sample frame construction, DSF, coverage

1. Introduction

Response rates for random-digit dial (RDD) telephone surveys have steadily declined over the past decades (Curtin et al., 2005). Consequently, survey research organizations have been investigating alternative survey designs to overcome growing shortcomings in RDD. In turning to an address-based sampling (ABS) frame, researchers are afforded more control in coverage and can potentially access hard-to-reach populations and reduce survey non-response by integrating a multi-mode approach (Iannacchione et al., 2003; Link et al., 2009; Link et al., 2008). One primary disadvantage to ABS is its cost as it has historically been difficult to obtain accurate and up-to-date address lists for target populations or target areas in some situations. The gold standard for ABS frame construction has been field listing until recently, a process wherein field staff enumerate and order each housing unit in a specified area. Listing can be costly and subject to timeliness issues in areas undergoing development.

Survey research organizations have investigated using versions of the U.S. Postal Service Delivery Sequence File (DSF) as a replacement for traditionally field listed address lists to reduce costs (Iannacchione et al., 2003; Kennel and Li, 2009). Research has shown that coverage offered by the DSF can be adequate to use in place of the traditional lists in urban and suburban areas and that it is often more efficient to provide a field lister the DSF for enhancement rather than creating a list from scratch in places where the DSF coverage is questionable (English et al, 2012). The benefits of using the DSF for ABS frames are numerous. For one, using the DSF is considerably more cost-effective than traditional field listing. Also, because the address data found on the DSF are "cleaner" than the data collected in the field, it is easier to match to other databases, such as phone databases or market research databases (Brick et al., 2011). Such data standardization improves the ability of researchers to conduct multi-mode surveys and flag certain households as being part of a target group. Matching addresses to telephone numbers can bring costs closer to the level of RDD surveys (Link et al. 2008). Additionally, DSF lists are maintained by the U.S. Postal Service and as a result, are updated frequently to reflect housing changes. The constant update increases the accuracy of the frame and removes the costs of having to return to the field to update the frame (O'Muircheartaigh et al., 2007).

DSF Geography

One drawback to using the DSF is that on the surface it lacks the geographic information necessary to link it to census geography. Because of the high costs related to in-person field interviewing, many ABS frame designs incorporate a multi-level sampling approach – first selecting a sample of geographic areas like census tracts or census block groups, then selecting a sample of housing units within the selected target areas. As opposed to a sample of 1,500 addresses randomly scattered across the entire United States, for example, the end result of a multi-level sampling approach might be 1,500 address clustered in groups of 15 across a selection of 100 nationally representative census tracts, greatly reducing the interviewer travel time and field costs of the study. The difficulty with this kind of ABS design, however, is that the DSF does not contain information on the precise location of each address. While the postal service maintains information related to city, state, zip, carrier route and walk sequence of each address, postal workers are not in the business of determining in which census block each address lies.

To overcome the limitations of DSF geography, researchers must append additional geographic information to the DSF in a process called "geocoding". Typically performed using commercial software, geocoding an address involves matching it to a database and imputing its location along a street using an address range. While highly accurate, especially in urban and suburban areas, the geocoding process is often an unacknowledged source of error in ABS frame design (Eckman and English, 2012).

Specifically, geocoding error can result in “over-coverage” or “under-coverage” in the frame. The purpose of this research is to perform an exploratory analysis to determine the correlates of geocoding error with the end goals of gaining a better understanding of the potential impact of geocoding error on bias and providing suggestions on how to enhance frame-making decisions.

2. Background

The fall of 2011 offered a unique opportunity to study geocoding error. Every ten years, after the release of the preliminary decennial census counts, NORC creates a new “National Frame” that is used to select representative samples for many national area-probability studies conducted by NORC, such as the General Social Survey (GSS) and the Survey of Consumer Finances (SCF). Of the 1,516 segments selected – either census tracts or census block groups depending on the housing unit density – the vast majority were determined as having adequate DSF coverage based on the ratio of DSF to census counts. However, 126 segments were identified has having problematic DSF coverage that needed to be enhanced in the field.

Unlike past frame construction that was conducted with paper-and-pencil, the 2010 National Frame field listing was conducted using a handheld device that allowed for several enhancements. Listers were now able to collect latitude and longitude coordinates for each housing unit using an integrated GPS and take photographs using an integrated digital camera. In addition, use of the devices gave listers the ability to perform real-time searches of the DSF extract for each segment while in the field. New data collected by the device as well as methodological changes brought about by the searchable DSF extract provided a new opportunity to assess the geocoding quality of the DSF in the segments where listing was required. In particular, we were now able to identify the DSF housing units that had incorrectly geocoded and try to model the conditions under which geocoding error was more likely to occur. What follows is a description of our analysis.

3. Data and Methods

Due to variations in the quality of listing due to lister skill and device performance, we did not analyze data from all 126 DSF-enhanced segments. Instead, 21 segments were selected for analysis based on the following criteria. First, we selected segments where the GPS was working for at least 90% of the addresses listed. Next, we eliminated segments where the initial DSF coverage relative to Census was very low – below 30% – to remove those without enough geocoded addresses to analyze. Finally, we made a selection from the remaining segments based on what we believed to be a fairly representative section of listing conditions – urban, suburban and rural. Our selection consisted of 21 segments containing 8,560 DSF lines that were enhanced in the field.

As a next step, we needed to identify the DSF lines that were incorrectly geocoded into the National Frame segments. Of the 8,560 DSF addresses provided to the listers for enhancement, 7,504 or 88% were confirmed in the field, leaving 1,056 or 12% unconfirmed. There are number of reasons why a DSF line might not be confirmed in the field. A housing unit may be difficult to find or, in rural areas, be located on a street that has multiple naming conventions. In addition, though we give explicit instructions to list vacant housing units, listers may exclude homes that appear vacant. Despite these issues, we argue that a large majority of DSF addresses that went unconfirmed did so because of geocoding error. That is, they were not located within the selected sample geography as indicated by the geocoding software.

Taking this approach means are modeling DSF over-coverage only, and while there is a parallel problem of under-coverage due to geocoding errors, under-coverage is beyond the scope of this paper. The segments in which we are conducting our analysis are a subset of the 126 segments that were initially identified as having low DSF coverage. In order to measure under-coverage due to geocoding error, we would have had to match addresses added by the field listers to the DSF that geocoded outside of the segment. This process was deemed beyond the scope of this paper which instead attempts to model the probability, based on context and address characteristics, that an address provided for enhancement was confirmed as not existing by our field staff.

For our analysis we used a generalized linear model with a logistic link function. To increase interpretability of results and decrease processing time, we conducted our model on a random sample of 4,000 addresses from the 8,560 addresses in the target area. The response variable was a binary flag indicating whether or not the address was confirmed in the field as described above. Our independent variables fell into two major categories, those that describe the address and those that describe the census block into which the address was initially geocoded. Address-level variables were derived directly from the DSF. In addition to basic address information, observations made by postal workers of housing units along their routes are stored in a variety of flags that accompany the DSF. The DSF flags we included in our first model were two binary variables, vacancy status and drop point flag, and one categorical variable, record type. The vacancy flag is self-explanatory – it indicates whether the postal worker believes the unit to be vacant. Record type describes the type of housing unit it is: a single family home, a high-rise, or a rural route. The drop point flag indicates whether it is a multi-unit building in which the units have no unit designation and share a common mail box (Dekker et al, 2012).

There may not be a strong theoretical connection between vacancy or drop point delivery and geocoding error. It stands to reason that neither the lack of inhabitants nor the possibility that a housing unit has been subdivided into multiple, undesignated units should influence whether or not the location of an address can be correctly imputed. As such, we chose to include these variables as controls for lister error. Listers often exclude addresses that appear vacant, despite receiving instructions to include them, and drop point addresses often contain hidden units that are difficult to find in the field (Dekker et al, 2012). Conversely, we hypothesize that there is a relationship between geocoding error and the type of residence. The easiest kind of address to geocode should be a single family home, while we expect that the geocoder should have much more difficulty with non-city style address and slightly more difficulty with high-rise addresses.

In addition to address-level variables, we included several variables that describe the characteristics of the blocks into which the address were initially geocoded. The most important of these variables is the DSF-to-Census ratio. As indicated earlier, this variable when calculated at the segment level is used to measure the DSF coverage and to determine whether it is advisable to use the DSF as a stand-alone frame. We can expect individual blocks to have a more unstable DSF-to-Census ratio and as a result be more of an indicator of geocoding error. For example, we hypothesize that a block with four times as many DSF housing units as housing units enumerated by the Census is likely to experience geocoding error. Because we do not believe this variable to behave linearly, we turned the DSF-to-Census ratio into a categorical variable measuring four ranges: 0 to 0.9, 0.9 to 1.25, 1.25 to 2 and 2 and above; we used 0.9 to 1.25 as reference category.

Other block-level variables measure the housing unit distribution and neighborhood characteristics of the block. We use the Census Type of Enumeration Area[1] (TEA) code to estimate whether the block is rural or urban. In addition, we include a flag that indicates whether the block is within the boundaries of the principal city of the CSA, or Combined Statistical Area, (e.g., Chicago is the principal city of Chicago CSA). To provide a measure of block topology, we added the total area of the block as well as a flag indicating whether the block is adjacent to a water feature. Creeks and streams form the borders of many blocks and make it difficult to determine correct address ranges. Housing unit density, percent multi-unit dwellings, percent occupied housing units and percent owned all measure the character of the housing stock of a particular block. To measure the potential impact of geocoding error on bias, our model initially included demographic variables such as median household income, median house value and race percentages.

4. Results

Modeling Over-coverage Error

Results from the initial model are presented in table 1, which demonstrate that many of the findings are significant and agree with our original hypotheses. The model fails the Hosmer-Lemeshow test indicating that the model demonstrates goodness of fit. Belonging to a block with urban characteristics as indicated by TEA code, the principal city flag and high housing unit density was associated with decreased probability of incorrect geocoding. Non-city style addresses were associated with an increase in the probability of geocoding error while belonging to a block with a high percentage of multi-unit buildings demonstrated the opposite. The results for the two DSF flags confirmed our expectations about interviewer error. Vacant housing units were more likely to not be confirmed, as were drop point addresses.

Perhaps the most interesting finding is the relationship of block-level DSF-to-Census ratio and geocoding error. As hypothesized, the model demonstrates that belonging to a block with a very high DSF-to-Census ratio (at or above 2) is associated with increased

[1] TEA Code is used by the Census Bureau to indicate if a block is sufficiently urban to enumerate via mail-out/mail, or if it requires in-person address updating and data collection.

probability that a housing unit is not confirmed in the field. Surprisingly, though, belonging to a block with a medium-high DSF-to-Census ratio (between 1.25 and 2.0) is associated with similarly high odds of geocoding error as belonging to a block with low DSF-to-Census ratio (less than 0.9). This finding may influence future frame building decision. Because a low DSF-to-Census ratio at the block level is generally associated with an increased probability of geocoding error, it is not safe to assume that the DSF-to-Census ratio is accurate below a certain threshold. For example, consider the situation where a hypothetical researcher were determining if a segment with a borderline DSF-to-Census ratio (say 91%) needed to be listed. If it were determined that several of the Census blocks that composed this segment had a less than adequate DSF-to-Census ratio, then it may not be safe to assume that all DSF lines in the segment geocoded into the correct areas. In fact, because a low DSF-to-Census ratio at the block level is associated with higher levels of geocoding error, the segment in question might have coverage below originally calculated rate of 91%. At the very least, this finding should encourage researchers to look at DSF coverage at different levels of geography – segment and block – when making listing decisions.

Table 1. Odds Ratios for Variables Predicting Over-coverage due to Geocoding Error

Variable or Predictor	Odds Ratio	95 % Confidence Interval	
DSF-to-Census Ratio <.9 (vs. .9 to 1.25)	2.249***	1.412	3.583
DSF-to-Census Ratio 1.25 to 2 (vs .9 to 1.25)	2.370**	1.349	4.164
DSF-to-Census Ratio ≥2 (vs. .9 to 1.25)	4.294***	2.627	7.019
Urban block (TEA Code)	0.573***	0.461	0.712
Drop Delivery Flag	9.087***	4.455	18.535
Record Type – High rise (vs. Single Family Home)	0.663	0.233	1.886
Record Type - Rural route (vs. Single Family Home)	94.523***	8.568	>999.999
Vacant Flag	4.302*	1.124	16.466
In Principal City	0.166***	0.065	0.422
Area in Sq. Miles (mean centered)	1.082***	1.055	1.109
HU Density (mean centered)	0.999***	0.999	0.999
Pct. Multi-Unit (mean centered)	0.279*	0.081	0.958

* p ≤ .05, ** p ≤ 0.01, *** p ≤ 0.001

Incorporating Spatial Effects

One of the underlying assumptions of the standard logistic regression is that the errors (residuals) are independently and identically distributed. However, the occurrence of over-coverage geocoding error might not be spatially independent. If the assumption of independent errors were violated, the standard logistic approach might not be sufficient to analyze the spatially-correlated data. Taking spatial effects into account when analyzing spatial data would result in better parameter estimates and enhance the overall goodness of fit of the model. Intuitively, we expect the existence of spatial dependence among spatial data analysis. The occurrence of over-coverage geocoding error for an address might be dependent on the occurrence of geocoding error of its neighboring addresses. Some possible explanations for the cause of spatial dependence include similar environmental circumstances such as road and water features.

In order to incorporate spatial effect into the previous model, we use Moran's I to detect possible existence of spatial effects, if any (Moran, 1950). If the spatial effects are present, the relationship between the binary response variable and the predicting variables will be modified. The Rook contiguity-based spatial weight is chosen in this paper for spatial weight matrix. The matrix, W, is an N by N matrix where N equals the number of addresses. We use the census tabulation block as the geography to calculate the contiguity matrix and assume that addresses in a tabulation block are considered as neighbors of the addresses in a neighboring tabulation block. After computing the spatial weight matrix for mailing addresses, we test the presence of spatial autocorrelation using Moran's I. Ranging between -1 and 1, Moran's I measures spatial autocorrelation with positive values indicating clustering, zero values indicating randomness and negative values indicating a non-random "checkerboard" distribution. For the distribution of mis-geocoded addresses from our analysis, the observed Moran's I value is low, 0.0281, but positive and highly significant: the p-value is Moran's I statistic is less than 0.00001. This result indicates the presence of spatial autocorrelation.

Table 2. Odds Ratios for Variables Including Spatial Autocovariate Predicting Over-coverage due to Geocoding Error

Variable or Predictor	Odds Ratio	95 % Confidence Interval	
DSF-to-Census Ratio <.9 (vs .9 to 1.25)	2.109**	1.445	3.172
DSF-to-Census Ratio 1.25 to 2 (vs .9 to 1.25)	2.297**	1.443	3.72
DSF-to-Census Ratio ≥2 (vs .9 to 1.25)	3.99***	2.673	6.134
Urban block (TEA Code)	0.584***	0.486	0.702
Drop Delivery Flag	9.38***	5.184	17.281
Record Type – High rise (vs. Single Family Home)	0.659	0.268	1.534
Record Type - Rural route (vs. Single Family Home)	94.711***	15.417	>999.999
Vacant Flag	4.168*	1.88	11.987
In Principal City	0.169***	0.07	0.345
Area in Sq. Miles (mean centered)	1.078***	1.056	1.101
HU Density (mean centered)	0.999***	0.999	0.999
Pct. Multi-Unit (mean centered)	0.291*	0.099	0.785
Spatial Autocovariate	1.017*	1.005	1.028
* p ≤ .05, ** p ≤ 0.01, *** p ≤ 0.001			

Based on the spatial statistics calculated, we extend the standard logistic regression model by adding an additional spatially related variable to account for the unrepresented spatial autocorrelation. This is achieved by adding an autocovariate to standard logistic regression to represent the relationship of neighboring response variables (Besag, 1971).

The results can be found in Table 2. We find that the odds ratio of the autocovariate in the spatial model equals 1.017 and is statistically significant ($p < 0.05$). Because the odds ratio is above 1, the model indicates that incorrectly geocoded addresses are clustered in space to some extent, which confirms our expectation that common environmental circumstances might influence the presence of geocoding error. Though the high rise flag is no longer significant in the second model, including the spatial autocovariate reduces the AIC from 2683.9 to 2680.5.

5. Discussion and Conclusion

The purpose of this research was to perform an exploratory analysis on the correlates of geocoding error in order to gain a better understanding of its impact on frame-building decisions. Initial findings indicate that there are covariates available that could be used or should be investigated for better predictions of potential problems related to geocoding error. Our two models demonstrate a significant, negative association between block-level characteristics of urbanicity and geocoding over-coverage. Blocks that have an urban TEA code designation, are located in the principal city of a metro area or have a higher percentage of multi-unit buildings are less likely to experience geocoding error. By the same token, larger block area, a characteristic of rural places, is positively associated with geocoding error. At the address level, the pattern holds; DSF records flagged as rural addresses are demonstrated to be related to higher levels of geocoding error, while DSF records flagged as high-rise buildings are demonstrated to be related to lower levels of geocoding error. The most important takeaway from this research is the positive relationship between low DSF-to-Census ratio and over-coverage due to geocoding error. If the DSF cannot account for more than 90% of the housing units in a block, results from our models indicate that there is an increased likelihood that the housing units associated with that block geocoded incorrectly. When making decisions about where to list, researchers should consider overall segment-level DSF coverage but also the DSF coverage of component blocks. This will allow for a more robust frame building process.

6. References

Besag, J. E. 1972. Nearest-neighbour systems and the auto-logistic model for binary data. *Journal of the Royal Statistical Society, Series B: Methodological*, 34:75-83.

Bilgen, Ipek, Ned English, and Lee Fiorio. 2012. “Coverage and Data Quality Association in Enhanced Address-Based Sample Frames”. *Proceedings of the Survey Research Methods Section, American Statistical Association.*

Brick, J. Michael, Douglas Williams, Jill M. Montaquila. 2011. Address-Based Sampling for Subpopulation Surveys. *Public Opinion Quarterly,* 75(3):409-428.

Curtin, Richard, Stanley Presser, and Eleanor Singer. 2005. Changes in Telephone Survey Nonresponse over the Past Quarter Century. *Public Opinion Quarterly,* 69(1):87-98.

Dekker, Katie, Ashley Amaya, Felicia LeClere, and Ned English. 2012. Unpacking the DSF in an Attempt to Better Reach the Drop Point Population. *Proceedings of the Survey Research Methods Section, American Statistical Association.*

Eckman, Stephanie and Ned English. 2012. Creating Housing Unit Frames from Address Databases: Geocoding Precision and Net Coverage Rates. *Field Methods.*

English, Ned, Colm O'Muircheartaigh, Katie Dekker, Ipek Bilgen, Lee Fiorio, Mark Clausen, and Tamara Brooks. 2012. Predicting when to Adopt Given Frame Construction Methods: Modeling Coverage and Cost Benefits. *Proceedings of the Survey Research Methods Section, American Statistical Association.*

Iannacchione, V.G., J.M. Staab, and D.T. Redden. 2003. Evaluating the use of residential mailing lists in a metropolitan household survey. *Public Opinion Quarterly*, 67(2):202-210.

Kennel, Timothy L., and Mei Li. 2009. "Content and Coverage Quality of a Commercial Address List as a National Sampling Frame for Household Surveys." *Proceeding of the Joint Statistical Meetings.*

Link, Michael W., Gail Daily, Charles D. Shuttles, Tracie L. Yancey, and H. Christine Bourquin. 2009. "Building a New Foundation: Transitioning to Address-Based Sampling After Nearly 30 Years of RDD". *Proceedings of the American Statistical Association, AAPOR* [CD ROM], Alexandria, VA: American Statistical Association.

Link, Michael W., Michael P. Battaglia, Martin R. Frankel, Larry Osborn, and Ali H. Mokdad. 2008. "A Comparison of Address-Based Sampling (ABS) Versus Random-Digit Dialing (RDD) For General Population Surveys. *Public Opinion Quarterly*, 72(1): 6-27.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1-2):17-23.

O'Muircheartaigh, Colm, Ned English, Michael Latterner, Stephanie Eckman, and Katie Dekker. 2009. "Modeling the Need for Traditional vs. Commercially-Available Address Listings for In-Person Surveys: Results from a National Validation of Addresses." *Proceeding at the Joint Statistical Meetings.*

O'Muircheartaigh, Colm, Edward English, and Stephanie Eckman. Predicting the Relative Quality of Alternative Sampling Frames. *Proceedings of the Joint Statistical Meetings.*2007.

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*.