Approaches to Modeling the Characteristics of Undeliverable-As-Addressed Addresses in the American Community Survey

Kristen Cyffka, Steven P. Hefter
U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233

Abstract

Although ordinary logistic regression is a widely-used tool, such models are often inappropriate given complicated data structures. We discuss methods to assess the quality of logistic regression models and explore alternatives to traditional regression models. To illustrate our findings, we investigate if characteristics of an address in the American Community Survey (ACS) can predict if a mailing is undeliverable as addressed (UAA) by the United States Postal Service. In 2012, local post offices reported that over 10% of mailed questionnaires in the ACS were UAA. By identifying the address and geographic characteristics for those mailings which are returned as UAA, we hope to identify certain types of addresses that are especially problematic and to provide suggestions for their improvement. To obtain this information, we will compare a variety of logistic regression approaches including mixed effects, generalized estimating equations, and spatial models. We will also investigate the use of classification trees for variable selection. We will discuss how to select an appropriate model and if our results can inform approaches to decrease the ACS UAA rate.

Key Words: American Community Survey, goodness of fit, model evaluation, logistic regression, classification trees

1. Introduction

Logistic regression is a popular tool to describe the characteristics of a population. At many survey organizations, it is often used to look at address frames or at responding households. However, many of these attributes are summary statistics collected at a higher level such as census tract or county. This could introduce within-cluster homogeneity, and a traditional logistic regression model may have correlated errors within clusters and thus biased tests.

We examine alternatives to traditional logistic regression and discuss why they may be more appropriate in dealing with household- and address-level data. To illustrate this discussion, we examine the characteristics of Undeliverable-As-Addressed (UAA) addresses in the American Community Survey (ACS).

The ACS mailed paper questionnaires to approximately 3.4 million addresses in 2012. Of these, 18 percent had mailings that were returned by local post offices because they

were UAA. The Census Bureau is investigating ways to decrease the rate of mailings that are returned UAA, which has the potential to decrease mailing and nonresponse-follow-up costs as well as the workload burden on the United States Postal Service (USPS). We specifically focus on identifying characteristics of UAA addresses of mailable households in the United States in the 2012 sample.

1.1 The American Community Survey

The ACS is a national survey that's collected monthly. That is, we draw a yearly stratified systematic sample which is divided into 12 panels. A new panel begins data collection each month. The ACS is used to provide estimates of the demographic, housing, and socio-economic characteristics of housing units and group quarters in the U.S. and Puerto Rico. In this paper we focus purely on housing units in the U.S. in the 2012 sample.

The addresses from these units come from a number of sources. In late 2009, census workers looked for every place where people could live or stay to update our address list. This was the Address Canvassing operation, which is part of the Decennial Census. This address list stemming from Decennial Census activities is continuously updated with the U.S. Postal Service's Delivery Sequence file, which is the set of mail delivery points used by the USPS. We also draw address tracking information from various other Census Bureau operations.

In 2012 and earlier, the ACS consisted of three data collection modes. Generally, we first mailed an address an initial paper questionnaire. If there was no update from the household or the USPS within about two weeks, we mailed a replacement paper questionnaire. If the address didn't respond by mail, we called the case for a telephone interview. If there was still no response, it was eligible to be sampled for a personal visit.

Thanks to a USPS program, we get a discounted postal rate if our addresses fulfill certain criteria. Because of these restrictions, we need to guarantee that an address is mailable. Thus, we do an initial screening to ensure the addresses are of sufficient quality, and only this subset of addresses is eligible for mail. Certain types of addresses are not eligible for mailout, as are addresses in ZIP codes that had a 100 percent UAA rate in 2006. Addresses that are mailable are sent using pre-sorted first-class mail.

1.2 Undeliverable-As-Addressed Mailings

Despite our best efforts, 18 percent of these 3.4 million mailable addresses had at least one questionnaire that was returned because it was UAA. In other words, 13 percent of the 5.6 million questionnaires mailed to the 3.4 million mailable US addresses in the ACS housing unit sample were UAA. This means a postal carrier identified an issue with the mailing, stamped it "Undeliverable As Addressed", and sent that information back to the Bureau's National Processing Center. The mailing was then marked UAA in our system, and we didn't try to mail a second questionnaire to that address.

There are many reasons an address could be UAA, not all of which are associated with the actual address. For instance, an address may be UAA because the housing unit is vacant, the respondent refuses the mailing, or there is no mail receptacle. A full list of the twenty types of UAA addresses is shown in Table 1. The ACS does not collect the reason an address is UAA because of the extra cost of this service from the USPS.

Table 1: Undeliverable-As-Addressed Reason Codes

Endorsement	**Reason for Nondelivery**
Attempted, Not Known	Delivery attempted, addressee not known at place of address.
Box Closed, No Order	Post office box closed for nonpayment of rent.
Deceased	Used only when known that addressee is deceased.
Delivery Suspended to Mail Receiving Agency	Failure to comply with 508.1.8.1 through 508.1.8.3.
Illegible	Address not readable.
In Dispute	Mail returned to sender by order of chief field counsel (or under 508.1.0 and 508.2.0) because of dispute about right to delivery of mail.
Insufficient Address	Mail without number, street, box number, route number, or geographical section of city or city and state omitted and correct address not known.
Moved, Left No Address	Addressee moved and filed no change-of-address order.
No Mail Receptacle	Addressee failed to provide a receptacle for receipt of mail.
No Such Number	Addressed to nonexistent number and correct number not known.
No Such Office in State	Addressed to nonexistent Post Office.
No Such Street	Addressed to nonexistent street and correct street not known.
Not Deliverable as Addressed, Unable to Forward	No change-of-address order on file; forwarding order expired.
Outside Delivery Limits	Addressed to location outside delivery limits of Post Office of address. Hold mail for out-of-bounds customers in general delivery for specified period.
Refused	Addressee refused to accept mail or pay postage charges on it.
Returned for Better Address	Mail of local origin incompletely addressed for distribution or delivery.
Returned for Postage	Mail without postage or indication that postage fell off.
Returned to Sender, Contains Nonmailable Contents.	Mail returned to sender due to contents that are nonmailable.
Returned to Sender Due to Addressee's Violation of Postal False Representation and Lottery Law	Mail returned to sender under false representation order and lottery order.
Returned to Sender Due to Addressee's Violation of Postal False Representation Law	Mail returned to sender under false representation order.
Returned to Sender Due to Addressee's Violation of Postal Lottery Law	Mail returned to sender under lottery order
Temporarily Away	Addressee temporarily away and period for holding mail expired.
Unclaimed	Addressee abandoned or failed to call for mail.
Undeliverable as Addressed, Missing PMB or # Sign	Failure to comply with 508.1.8.2e..
Vacant	House, apartment, office, or building not occupied. (Use only if mail addressed “Occupant.")

Source: USPS, http://pe.usps.com/text/dmm300/507.htm

The Census Bureau is interested in methods to decrease this UAA rate. To do this, we examine the types of addresses that are UAA, which involves finding a best model for our rich but very unbalanced data. Using the characteristics illuminated by this model, we investigate if we could suggest a method to decrease the rate of undelivered mailings. Finding a way to improve addresses would improve the efficiency of our data collection. For instance, it would decrease workloads for the subsequent telephone and personal visit modes. Ideally, we'd be able to find low-quality addresses that have a very high probability of being UAA. We could find a way to improve the address so that they'll receive their mailings. If that's not possible, we'd investigate suppressing mailout to these pre-identified addresses and then moving them straight to later data collection modes. In this paper, we focus on finding the most useful model to approach this problem.

1.3 Alternatives to Logistic Regression

Ordinary logistic regression assumes that observations are independent (McCullagh, 1989). However, addresses are inherently geographic: we tend to expect addresses that are close to one another to have similar characteristics. Furthermore, several potential predictors are summarized at higher levels of geography, which inherently groups address within that geographic area. In these situations, errors are not independent, and we require and extension of ordinary logistic regression to provide unbiased tests. The decision of whether an address is UAA is made at the local post office level, which raises concern about the independence of reporting UAA information within ZIP code.

In environmental problems, there's evidence that introducing spatial random effects may improve fit. With this in mind, we developed these three models to extend generalized linear regression using available spatial information.

We first examine a generalized linear mixed-effects model (Affifi, 2004), which uses maximum likelihood estimation and accounts for potential clustering within ZIP code using random effects. Next we look at a population average model, which uses generalized estimating equations (GEEs) and a quasi-likelihood approach (Hubbard, 2010). Like the generalized linear mixed-effects model, it also allows for similarities within ZIP code. Finally, we use a spatial regression model that controls for spatial dependence by using the coordinates of each address (Ward, 2008). This allows us to acknowledge that a unit is related to its neighbors. We compare the results of each approach to see if results change if we adjust the underlying assumptions. We also investigate if it's worthwhile to move beyond a standard generalized linear model. Because we always want to maximize simplicity, we want to be sure that it would really be necessary to use one of these more involved methods.

2. Methodology

2.1 Data Sources

Data from this project comes from several Census Bureau and USPS files. We use the ACS 2012 housing unit second-stage sample file and the ACS 2012 housing unit sample delivery file, which contain information used to create mailing labels; the August 2011 USPS ZIP+4 file, which has carrier route information; the ACS control file, which tracks the status of every case in the ACS; and the sampling frame, the ACS 2012 edited master address file extract, which includes information from the Census Bureau Geography

Division's master address file, the USPS Spring 2011 DSF, and decennial census operations that have contacted that address (U.S. Census Bureau, 2012). As always, these data are subject to error arising from a variety of sources. For each address, we have over 300 independent variables. Most of these predictors are categorical, and many of them have small counts for some levels. These small cell sizes introduce optimization and computing constraints into logistic regression procedures.

2.2 Analyses

We restrict our analyses to those 3.4 million addresses in the U.S. that were mailed questionnaires in 2012. The dependent variable is whether an address had at least one questionnaire returned UAA. Because the two mailings are sent only about two weeks apart, some UAA addresses end up getting both mailings because we didn't find out in time. Out of those 3.4 million addresses sent a mailing in 2012, about 447,000 had only the first mailing returned UAA. About 107,000 had both mailings returned UAA, and 65 thousand had only the second mailing returned UAA (Table 2).

Table 2: 2012 ACS Undeliverable-As-Addressed Mailings

UAA Mailing	Mailable Addresses	
Total	3,436,723	100%
Neither questionnaire	2,817,041	82%
Initial only	447,003	13%
Replacement only	65,366	2%
Both questionnaires	107,313	3%

We first use a classification tree (Breiman, 1984) to perform variable selection on the 300 potential predictors. Because of computing constraints, we built a series of classification trees with random subsamples using R's rpart package (Terry, 2006). Going forward, we restricted analyses to the 94% of mailable addresses that had a latitude and longitude, where addresses with and without coordinates had similar UAA rates. A traditional logistic regression model used these selected variables to predict if an address is UAA by using the GENMOD procedure in SAS® (SAS® Institute Inc., 2008). Next we fit a logistic mixed-effects model with a random effect for ZIP code using the GLIMMIX procedure in SAS® (SAS® Institute Inc., 2008). We also cluster by ZIP code using a GEE logistic model, again using GENMOD. We investigate Moran's I and a visual display of the geography of UAA addresses, then run a spatial regression model using latitude and longitude. These use the VARIOGRAM (SAS® Institute Inc., 2008) and GLIMMIX procedures in SAS®. Finally, we compare the results of these methods and discuss the advantages and disadvantages to using each. We consider the types of addresses this research predicts are UAA, and we provide suggestions for future research into finding actionable ways to decrease the UAA rate.

3. Results

3.1 Variable Selection

Using classification trees, we chose seven variables:

- whether an address was excluded or included in USPS delivery statistics;
- the DSF record type, which tells if the record refers to a high-rise apartment, a "street" unit, or one on a rural route;

- the ACS address flag, which is used to pinpoint addresses we expect to be troublesome based on USPS information;
- whether the address was on the spring 2011 DSF, which was the most recent version used for these addresses;
- the number of units at the basic street address;
- the address characteristic type code for the census block, which tells the prevalence of city-style addresses in the block. A city-style address is something like "101 Main Street, Anytown, State, USA". Another type of address is a rural route, which would simply say "Rural Route 1, State, USA"; and finally
- the mailing and location address match score, which is the number of matching mailing and location address components.

Because some of these variables had many levels, we collapsed some categories to get more evenly-distributed cell sizes. To do this, we used information from how the classification tree split variables into groups more and less likely to be UAA. For a description of these final edited variables, see Table 3.

3.2 Modeling

First we fit the traditional logistic model, to predict if an address had at least one questionnaire stamped UAA using our seven main descriptive variables.

There was evidence that each of the three extensions was an improvement on the traditional logistic model. In the mixed-effects model, the random effect for ZIP code was small but non-zero, which means that it has some effect in our model. Thus, it's a meaningful term, and this model is an improvement on the traditional model. The GEE model was also an improvement on the traditional model. To evaluate this, we use Pan's Quasilikelihood under the Independence Model Criterion (QIC). For the spatial problem, we first looked at Moran's I, which describes if a spatially-oriented variable is clustered, randomly distributed, or dispersed. In our case, it was significantly different than zero ($p<0.0001$) and positive, which points to some clustering effect. To fit the regression model, we used a spherical variogram, which was the closest approximation to our empirical variogram.

To compare how powerful the models were at classifying UAA and non-UAA cases, we looked at the receiving operation characteristic (ROC) curves (Figure 1). ROC curves plot the model's sensitivity, or rate of true positives, against 1 less its specificity, or the rate of false positives. This describes how well the model can classify positive and negative cases. A better ROC curve hugs the top left corner of the graph and so has area closer to one. The mixed-effects model, GEE, and spatial models all had an area in the low 80 percent range. The mixed effects model was slightly better than the other two: its area was 85 percent. The GEE and spatial models were not far off at about 83 percent. The bumpiness in the mixed effects' ROC curve may explain this difference. These areas are not bad, but because we're hoping to specifically identify UAA addresses with a high level of certainty, this may not be good enough for our purposes. The models may not be able to perform well because we're trying to predict such a wide range of outcomes: a vacant unit may not be the same as one with a poor address.

Table 3: Variables of Interest from the 2012 ACS Mailable Housing Unit Sample

Variable	**Frequency**	**(%)**	**% UAA**	**χ^2 Test**
Total	3,436,723	(100)	18	-
Undeliverable As Addressed				
Yes	619,682	(18)	100	
No	2,817,041	(82)	0	
Excluded from USPS Delivery Statistics				p<0.0001
Excluded	578,348	(17)	62	
Included	2,858,375	(83)	9	
DSF Record Type				p<0.0001
High-rise apartment, office building, or firm	604,314	(18)	15	
Street, rural route, or highway contact	2,384,356	(69)	11	
Blank	448,053	(13)	61	
ACS Address Flag				p<0.0001
Not a DSF "X record"	3,291,108	(96)	16	
Multi-unit placeholder or ZIP code conversion	3,971	(0)	49	
Unknown	141,644	(4)	61	
On the Spring 2011 DSF				p<0.0001
On the DSF	2,820,329	(82)	9	
Not on the DSF	616,394	(18)	59	
Number of Units at the Basic Street Address				p<0.0001
One unit	2,750,590	(80)	18	
More than one unit	686,133	(20)	17	
Block Address Characteristic Type Code				p<0.0001
Non-residential or PO box	8,714	(0)	23	
City-style	2,619,599	(76)	13	
City-style and noncity-style	723,906	(21)	34	
Assorted noncity-style	12,609	(0)	48	
Rural route	3,400	(0)	33	
Blank or no addresses found	68,495	(2)	25	

Variable	**Mean (SD)**		***t* Test**
	Not UAA	**UAA**	
Mailing and Location Address Match Score	16.6 (0.9)	16.4 (1.2)	p<0.0001

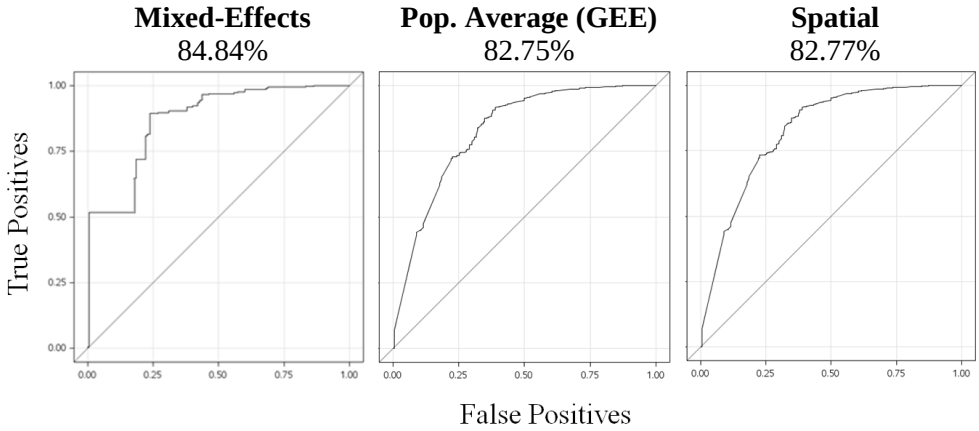

Figure 1: Receiving Operator Characteristic Curves for the 3 Extended Models

However, aside from how these models perform, it's also worth considering the costs and benefits of each approach. Of course, in the words of George Box, "all models are wrong, and some models are useful". We'd like to determine which is the most useful for our problem.

As mentioned before, the traditional logistic model has some independence assumptions that we just can't meet. However simple it is, it doesn't do our data justice. The mixed-effects model, like the standard generalized linear model, is rather more sensitive to its assumptions. However, the GEE model is not as sensitive: even if our selected correlation structure isn't correct or isn't complete, the GEE model will still perform well. This is a definite advantage of the GEE model. On the other hand, the mixed-effects model can give estimates of the impact of the different random effects, which can be useful; however, this isn't particularly important to us. The spatial model seems like a great choice for looking at whether these addresses are clustered, and it can account for clustering between ZIP codes. In Figure 2, we see evidence of clustering between ZIP codes and not just within ZIP codes. Spatial analysis does require coordinates for every data point, which are not always available, and it's computationally intensive: it can't handle the size of our data. The mixed-effects model had similar computing problems, though not as bad as the spatial. The GEE model had no such issues. This and the robustness when faced with misspecification make the GEE approach a very practical and attractive option, though given more computing power we would prefer to use the spatial model.

All four models produced the same conclusions of significance for each variable, so they all have the same basic results.

3.3 Subject Matter Conclusions

Table 4 shows the results of the GEE model that allows for clustering within ZIP code. The mailing and location address match score was not significant in the final model, and so was removed following a backwards elimination procedure using a significance level of 0.01. This model indicates that after controlling for the rest of the variables:

- Addresses that are excluded from USPS delivery statistics are more likely to be UAA than those that are included.
- Addresses of high-rise apartments, office buildings, or firms or those that have no DSF record type are more likely to be UAA than addresses on streets, rural routes, or with a highway contact.
- Addresses that have a multi-unit placeholder, have had a ZIP code conversion, or have no ACS address flag are more likely to be UAA than those that are not an X record on the DSF.
- Addresses not on the most recent DSF are more likely to be UAA than those on the most recent version.
- Addresses that have only one unit at the basic street address are more likely to be UAA than basic street addresses that have more than one unit.
- Though blocks with entirely city-style and assorted noncity-style addresses do not have different UAA rates, and blocks with mixed city-style and noncity-style addresses are more likely to be UAA than blocks with all city-style addresses.
- Addresses in a entirely city-style block are less likely to be UAA than addresses on blocks that are entirely non-residential, entirely PO box, entirely rural route, no addresses found, or with a blank descriptor.

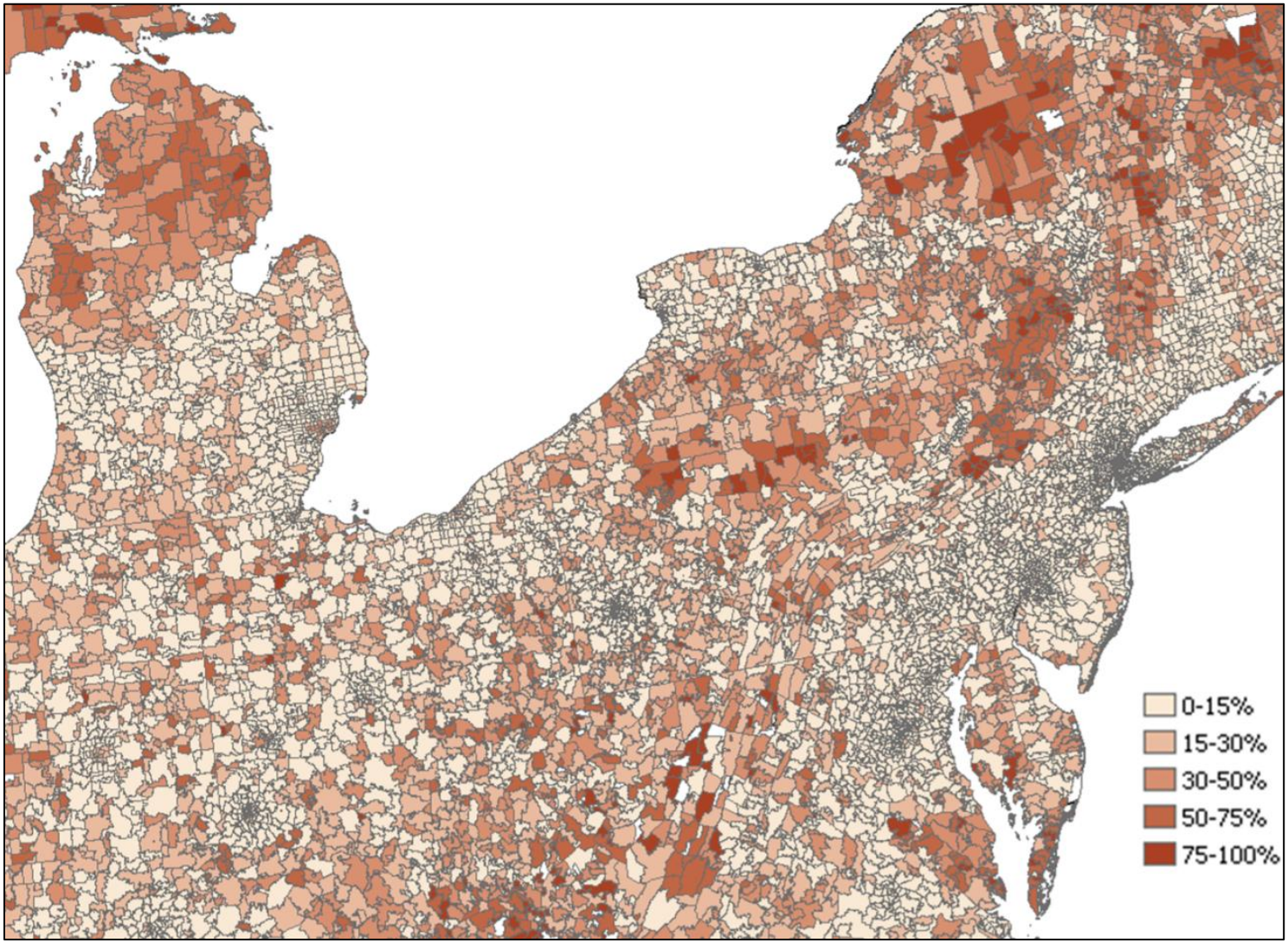

Figure 2: 2012 ACS UAA Rates by ZIP Code in the Northeast United States

Table 4: Results of the Generalized Estimating Equations Model

Variable	Estimate	(SE)	P>|Z|
Constant	-2.3742	(0.0071)	-
Excluded from USPS Delivery Statistics			
Excluded	0.9591	(0.0206)	p<0.0001
Included			
DSF Record Type			
High-rise apartment, office building, or firm	1.0256	(0.0104)	p<0.0001
Street, rural route, or highway contact			
Blank	0.9732	(0.0300)	p<0.0001
ACS Address Flag			
Not a DSF "X record"			
Multi-unit placeholder or ZIP code conversion	0.7132	(0.0463)	p<0.0001
Unknown	0.9361	(0.0303)	p<0.0001
On the Spring 2011 DSF			
On the DSF			
Not on the DSF	0.7173	(0.0186)	p<0.0001
Number of Units at the Basic Street Address			
One unit			
More than one unit	-0.4477	(0.0109)	p<0.0001
Block Address Characteristic Type Code			
Non-residential or PO box	-0.2244	(0.0411)	p<0.0001
City-style			
City-style and noncity-style	0.2417	(0.0066)	p<0.0001
Assorted noncity-style	0.0444	(0.0300)	p=0.1389
Rural route	-0.4141	(0.0600)	p<0.0001
Blank or no addresses found	-0.0523	(0.0170)	p=0.0021

Generalized linear model using a logistic link function and a random effect for ZIP code

3. Discussion

There are several drawbacks to keep in mind when considering this research. The classification tree was very useful in selecting variables, but there is some concern the variables may not behave the same way in a regression model as they do in a classification tree. It is a straightforward and intuitive method to use; however, other more sophisticated methods of selection may yield better results.

This work demonstrates that using more sophisticated models is worthwhile. Significance testing is not reliable when there are spatially-oriented errors, which is likely common in much of the work done with the Census Bureau's heavily-geographic data. In our case, controlling just for ZIP code is a rather naïve approach. As demonstrated in Figure 2, there are regional differences in UAA rates that go beyond the ZIP code. Given more computing power, the spatial analysis would better control for this clustering effect.

A large underlying issue with this analysis is that a mailing being returned UAA is not be a good indication of a problem with the address. Though the Census Bureau would like to decrease the ACS UAA rate, this should likely be focused on addresses whose written address needs improvement. It may not be worthwhile to suppress mailout to vacant units, which status could change in the year between updating the predictors and mailout.

A similar concern exists for other units with temporary features like those with no mail receptacle. However, there is research in Decennial Census operations that suggests a UAA reason of vacancy may be a good predictor of a vacant unit (Jackson, 2013), though this method may undercount traditionally under-represented groups. That Decennial Census study also found that the majority of UAA addresses were vacant. In the ACS, the only study into these UAA reason codes was done in one panel in Puerto Rico in 2006. This showed that 32% of UAA addresses were undeliverable because of issues unrelated to address quality. Though we know that mailing issues in Puerto Rico are very different than those stateside, these two sources still suggest that we may not want to expect all UAA addresses to have the same characteristics, especially when we are only most concerned with the quality of the written address.

Based on this work, we shouldn't draw conclusions about address quality. Of course, we could still explore potential adaptive-design-type updates to the current contact methodology; for example, if we could confidently predict that an address is very unlikely to receive mailings, then it may not be worthwhile to mail it a questionnaire in the first place. However, our models aren't very precise because they're attempting to predict such a wide range of outcomes as shown in Table 1, and we can't very specifically predict which addresses will have a mailing returned UAA.

In the future, it would be valuable to collect UAA reason in the ACS. Even codes for just one panel in the United States would give us an idea of vacancy rates and the magnitude of the problem with poor-quality addresses. Using this information as well as the wealth of other data available at the Census Bureau, we would finally be able to address this large mailing issue, investigate ways to improve poor-quality addresses, and ultimately improve the efficiency of our data collection strategies.

Acknowledgements

Thank you to Lawrence Bates, Deborah Griffin, Geoffrey Jackson, Daniel Sommers, Donald Keathly, and Star Ying for your valuable insight.

References

Affifi, Abelmonem, Virginia A. Clark, and Susanne May. (2004). *Computer-Aided Multivariate Analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Breiman, Leo, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. (1984). *Classification and Regression Trees*. Boca Raton, FL: Chapman & Hall/CRC.

Hubbard, Alan E. et al. (2010). "To GEE or Not to Gee: Comparing Population Average and Mixed Models for Estimating the Associations Between Neighborhood Risk Factors and Health." *Epidemiology*. 21.4: 467-474.

Jackson, Geoffrey I. (2013). "Identifying Vacant Housing Units in the 2010 Census with UAA and Pre-enumeration Information." U.S. Census Bureau.

McCullagh, P., and J.A. Nelder. (1989). *Generalized Linear Models.* Boca Raton, FL: Chapman & Hall/CRC.

SAS® Institute Inc. (2008). "The GENMOD Procedure." *SAS/STAT® 9.2 User's Guide*. Cary, NC: SAS Institute Inc.

SAS® Institute Inc. (2008). "The GLIMMIX Procedure." *SAS/STAT® 9.2 User's Guide*. Cary, NC: SAS Institute Inc.

SAS® Institute Inc. (2008). "The VARIOGRAM Procedure." *SAS/STAT® 9.2 User's Guide*. Cary, NC: SAS Institute Inc.

Terry M Therneau and Beth Atkinson. R port by Brian Ripley. (2006). rpart: Recursive Partitioning. R package version 3.1-32.

U.S. Census Bureau. (2012). "American Community Survey Design and Methodology Report."

Ward, Michael D., and Kristian Skrede Gleditsch. (2008). "Spatial Regression Models." *Quantitative Applications in the Social Sciences*. Thousand Oaks, CA: SAGE Publications.