UNITED STATES DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. Census Bureau
Washington, DC 20233-0001

September 27, 2012

2012 AMERICAN COMMUNITY SURVEY RESEARCH AND EVALUATION REPORT MEMORANDUM SERIES #ACS12-RER-29

DSSD 2012 AMERICAN COMMUNITY SURVEY MEMORANDUM SERIES #ACS12-CAUS-08

MEMORANDUM FOR ACS Research and Evaluation Steering Committee

From: Anthony G. Tersine ***/Signed/***
Assistant Division Chief, American Community Survey Methods Area
Decennial Statistical Studies Division

Prepared by: Bryan Schar
Community Address Updating System Branch
Decennial Statistical Studies Division

Subject: Refreshing the Community Address Updating System Block Score for 2011 Selection

Attached is the report for the American Community Survey Research and Evaluation project "Refreshing the Community Address Updating System Block Score for 2011 Selection." This project analyzed the post 2010 Decennial Address Canvassing results of the Demographic Area Address Listing program to update the selection score that the Community Address Updating System branch used for its March 2011 through February 2012 targeted block listings.

If you have any questions about this report, please contact Bryan Schar (301) 763-8498 or Anthony G. Tersine at (301) 763-1994.

Attachment

cc:
ACS Research and Evaluation Team
Deborah Griffin (ACSO)
Patrick Joyce (CRSM)
Freddie Navarro (DSSD)
Steve Hefter

American Community Survey Research and Evaluation Program

9/27/2012

Refreshing the Community Address Updating System Block Score for 2011 Selection

FINAL REPORT

Bryan Schar
Decennial Statistical Studies Division

Executive Summary

This project analyzed the results of Demographic Area Address Listing field operations conducted between April 2010 and January 2011 to update the block selection score that the Community Address Updating System branch used for its March 2011 through February 2012 targeted block listings. Specifically, the following question was considered:

Based on recent listing results, what linear combination of its three block scores, or factors derived from these scores, should the Community Address Updating System branch use when selecting blocks to list for March 2011 through February 2012?

The scope of this project was restricted to linear combinations of the three block scores to ensure actionable results were available by the time blocks were to be selected.

Background

Since September of 2010 the Community Address Updating System branch has used information from three predictive models when assigning selection scores to blocks. The models are based on the following methods:

- Decision Tree Learning
- Neural Network Analysis
- Regression Analysis

Each model is used to assign a score to each block. This score provides a measure of the predicted number of new addresses that will be added to the Master Address File by listing a block, with higher block scores being associated with more adds.

The score used to select to blocks for listing in September 2010 through February 2011 was the simple average of the three models' scores. Thus the selection score for block *i* (S_i) was:

$$S_i = 0.33 \cdot D_i + 0.33 \cdot N_i + 0.33 \cdot R_i, \qquad \text{(ES.1)}$$

where

D_i = decision tree based score for block *i*,
N_i = neural network based score for block *i*,
R_i = regression based score for block *i*.

When selecting blocks for listings, preference is given to blocks with higher selection scores.

Project Stages

This project was conducted in two stages. The results of the initial stage were used when selecting blocks that would be listed in March 2011 through August 2011. The results of the final stage were used when selecting blocks that would be listed in September 2011 through February 2012.

Data Used

Different vintages of Demographic Area Address Listing results data were used in each stage of this project. Table ES1 gives the vintages of data used in each stage.

Table ES1: Vintages of Data Used in each Project Stage

Stage	**Data Vintage(s)**
Initial	April 2010 - September 2010 (Preliminary Results)
Final	April 2010 - August 2010 (Final Results) September 2010 - January 2010 (Preliminary Results)*

* Only blocks without adds considered

Note that the preliminary results of a listing reflect data received from field representatives (FRs) that have not yet been vetted for inclusion into the Master Address File. The final results reflect the actual changes to the Master Address File due to the listing.

Results

Linear combination were constructed using regression analysis with the adds per block in the Demographic Area Address Listing results as the dependent variable, and the models' scores for blocks as the independent variables. These combinations were then compared based on their ability to identify the blocks with the most adds, with secondary comparisons conducted on their ability to identify blocks with high adds per preexisting unit and blocks with at least one add. Based on an analysis of these data, the linear combinations of the three block scores given in Table ES2 were produced.

Table ES2: Linear Combination Produced by each Project Stage

Stage	**Linear Combination Produced**	**Listing Period Applied**
Initial	$S_i = 5.92 \cdot ND_i + 7.55 \cdot NN_i - 6.54 \cdot NR_i$	March 2011 - August 2011
Final	$S_i = 4.99 \cdot ND_i + 4.12 \cdot NN_i - 3.03 \cdot NR_i$	September 2011 - February 2012

Sources: U.S. Census Bureau, Community Address Updating System Database, September 2010; U.S. Census Bureau, Master Address File Updating Files, April 2010 – January 2011; U.S. Census Bureau, Master Address File Transaction file, January 2011.

where

ND_i = normalized decision tree based score for block i,
NN_i = normalized neural network based score for block i,
NR_i = normalized regression based score for block i.

Model scores were normalized using the following method:

$$normalized\ score = \frac{score - \min(score)}{\max\ score\ - \min(score)} \qquad \text{(ES.2)}$$

1. Background

The Master Address File (MAF) is a U.S. Census Bureau file that contains an inventory of all known living quarters in the United States. It is the sole source of housing unit records for the American Community Survey's sampling frame, and the main source of group quarters (GQ) information for the ACS's GQ sampling frame (Bates, 2011). Between Decennial Censuses, the U.S. Postal Service's Delivery Sequence File (DSF) is the primary source of city-style[1] address updates for the MAF. The Demographic Area Address Listing (DAAL) operation, of which the Community Address Updating System (CAUS) targeted census block listing operation is a part, is the primary source of non-city-style[2] address updates (GEO, 2011b).

The CAUS branch is responsible for selecting the blocks that are listed through the CAUS program. Since September of 2010, the CAUS branch has used information from three predictive models when selecting blocks. The models are based on the following methods:

- Decision Tree Learning
- Neural Network Analysis
- Regression Analysis

See Bishop (1995), Larose (2005), SAS Institute Inc. (2003), and Tan, Steinbach, and Kumar (2006) for more information about these modeling methods.

Each model is used to assign a score to each block. This score provides a measure of the number of new addresses that the model predicts will be added to the Master Address File by listing a block, with higher block scores being associated with more adds. Attachment A provides a brief summary of how these models were constructed.

The selection score assigned to blocks listed from September 2010 through February 2011 was the simple average of the three model scores. Thus the selection score for block i (S_i) was:

$$S_i = 0.33 \cdot D_i + 0.33 \cdot N_i + 0.33 \cdot R_i, \tag{1.1}$$

where

D_i = decision tree based score for block i,
N_i = neural network based score for block i,
R_i = regression based score for block i.

An equal factor was assigned to each score based on the null hypothesis that the models were equally effective in identifying blocks with adds.

[1] Addresses that contain at least a house number and street name.
[2] Examples of non-city-style addresses are PO Boxes, rural routes, and description only addresses.

2. Research Goal

The goal of this research project was to use results of DAAL field operations conducted between April 2010 and January 2011 to answer the following question:

Based on recent listing results, what linear combination of its three block scores, or factors derived from these scores, should the Community Address Updating System branch use when selecting blocks to list for March 2011 through February 2012?

The main criteria used to answer this question was how well the top three to five percent of blocks as ranked by potential linear combination compared, in terms of total adds, to the top three to five percent of blocks as ranked by actual adds. This was done because funding at the time of the study only permitted about four percent of the blocks in the CAUS universe to be listed per year. The average adds per prelisting units and percent of blocks that had at least one add were examined.

By November of 2010, the CAUS branch had received the preliminary results for DAAL blocks listed from April 2010 through September 2010. These were used as the basis for determining whether the use of an alternative linear combination of the models' scores was justified for selecting the blocks to be listed during March 2011 through August 2011.

By February of 2011 the CAUS branch had received the final results of the DAAL blocks listed during April 2010 through August 2010, in addition to the preliminary results of DAAL blocks listed during September 2010 through January 2011. These data were used as the basis for determining whether to update the linear combination of the models' scores used to select the block to be listed during September 2011 through February 2012.

The CAUS branch restricted the scope of its research to linear combinations to ensure it had actionable results available within the limited timeframe it had to select blocks for listing.

3. Data Used

This research was conducted in two stages. The first stage was conducted in December of 2010 using the preliminary results of DAAL listings completed during April 2010 through September 2010. These listing results contained data from both the CAUS and current surveys programs, and were obtained from the Master Address File Updating Files (MAFUFs). These files contain listing results received from the field that have not yet been vetted by the Geography Division (GEO) for inclusion into the MAF.

The second stage of the research was conducted in February of 2011 using both the final results of DAAL listings completed during April 2010 through August 2010 and some of the preliminary results of DAAL listing conducted during September 2010 through January 2011. All final results were included in this stage of the research, but only those blocks whose preliminary results indicated they had no adds, and thus whose results would not be affected by the GEO's vetting process, were included. The preliminary data were obtained from the

MAFUFs, and the final data were obtained from the Master Address File Transaction File (MAFTA). The MAFTA reflected the actual changes to the MAF due to DAAL listings.

3.1. Initial Filter

Only a select subset of the above DAAL listing results was used in this research. Of the blocks for which there were results, only those blocks the met the following conditions were considered:

- The block contained 100 or less prelisting housing units (HUs). This filter was applied since, due to cost and time constraints, the CAUS does not currently list blocks with more than 100 prelisting HUs on the MAF.
- The Address Characteristic Type code of the block indicated it was from the traditional CAUS universe[3].
- The outcome flag for the block indicated it had been completely, as opposed to partially, listed.

There were 11,509 such blocks available for the initial stage of this research and 11,965 for the final stage. Within these blocks, only those preliminary adds that met the following conditions were considered in the initial stage of this research:

- The group quarters/housing unit and residential status flags on the MAFUFs indicated the added address was a residential housing unit.
- The unit status flag on the MAFUFs indicated that the address was either:
  - A valid address
  - A provisional add
  - Under construction
  - An empty mobile home/trailer site

Only those final adds that met the following conditions were considered in the final stage of this research:

- The Master Address File (MAF) update flag on the MAFTA indicated a new address was added to the MAF.
- The group quarters/housing unit flag on the MAFTA indicated the new address was a housing unit, and its MAFUF residential status flag indicated it was residential.
- The official unit status flag on the MAFTA indicated that the address was either:
  - A valid address
  - A provisional add
  - Under construction
  - An empty mobile home/trailer site

[3] See Attachments B and C for more information on the traditional CAUS universe and Address Characteristic Type codes.

3.2. Removal of Suspect Blocks (Initial Stage)

The CAUS program uses block listings conducted by Census FRs to improve the coverage of the MAF. When listing a block, FRs look for every place where people live, stay, or could live or stay. They compared what they see on the ground to what is shown on the MAF. Based on their observations, they verify, update, or delete addresses already on the MAF, and add addresses that are missing from it. However, since FRs are only human, they sometimes make mistakes. Common mistakes, in terms of inflating the adds in a block, are:

- Adding addresses that already exist on the MAF but are not geocoded to the block the field representative is listing.
- Adding addresses that the GEO determines actually represent duplicates of or changes to existing MAF records.
- Adding new addresses to the MAF, but geocoding them to the wrong block.

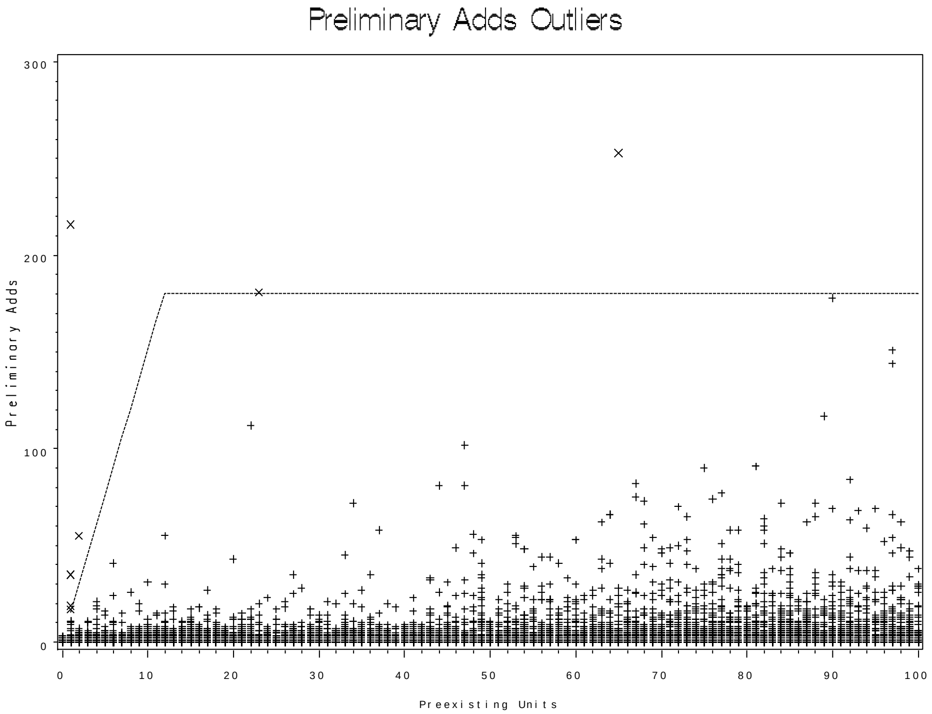
To maintain the quality of the MAF, the GEO vets the results of CAUS listings before including them on the MAF. A new MAF record is created in a block only when the GEO decides an add truly reflects a new unit on the ground in that block.

To try to reduce the number of "false" adds included in this research, an additional analysis was conducted on the preliminary results data to identify blocks that had an unusually large number of preliminary adds given the number of prelisting HUs they contained. The CAUS branch suspected that many of the adds in these blocks were "false" adds, and would be rejected by the GEO for inclusion on the MAF. The reason for this suspicion was that previous experience suggested that the number of adds and the number of prelisting HUs in a block were moderately positively correlated (Lawrence, 2010).

After exploring the data, the CAUS branch decided to exclude 8 blocks out of the total 11,509 that had more than 180 adds or more than 15 adds per prelisting units (when prelisting units was non-zero) from its initial analysis. These cutoffs were chosen heuristically by examining the distributions of the total adds, total prelisting HUs, and adds per HU for all the blocks. This analysis was not conducted during the final stage of this research, as it only considered those adds accepted by the GEO for inclusion on the MAF or blocks where no adds were found in the field.

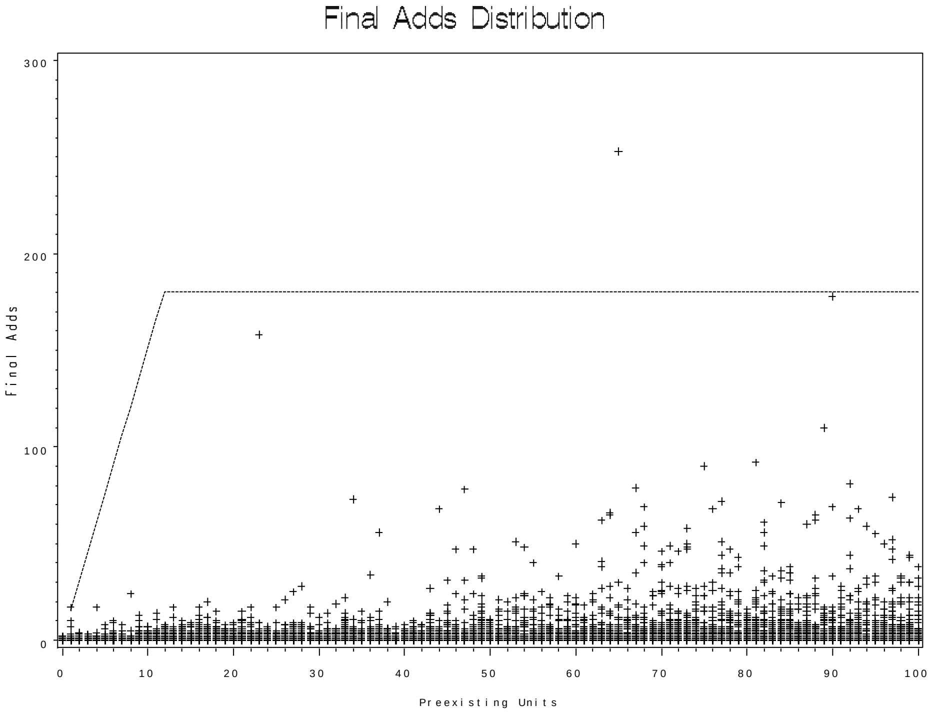
Figure 1 gives a plot of the distribution of preliminary adds vs. prelisting units per block. Blocks whose values fell above the line shown on this plot were excluded from the initial analysis. Figure 2 gives a plot of the distribution of final adds vs. prelisting units per block. The outlier decision line used for the initial stage of the research is included in this plot for comparison.

Figure 1: Preliminary Adds vs. Prelisting Units with Outlier Decision Line



Sources: U.S. Census Bureau, Master Address File Updating Files, April 2010 – September 2010.

Figure 2: Final Adds vs. Prelisting Units



Sources: U.S. Census Bureau, Master Address File Transaction File, January 2011; U.S. Census Bureau, Master Address File Updating Files, September 2010 – January 2011.

4. Methods Used

4.1. Linear Combination Creation

The REG procedure in SAS/STAT® software was used to produce the coefficients of the linear combinations developed during both the initial and final stages of this research. This procedure implements traditional least squares regression with the adds per block as the dependent variable and the three model scores as the independent variables. Both raw and normalized model scores were considered. Model scores were normalized using the following method:

$$normalized\ score = \frac{score - \min(score)}{\max\ score\ - \min(score)} \tag{4.1}$$

Normalization of the model scores was done so that all the models' scores fell within the same range. This made it easier to compare the different models' ability to identify and rank the blocks with the most adds without losing the "distance" information of the model scores (as opposed to a straight ranking). Though the models were all developed using the same data, their range of scores varied. This can be seen in Table 1, which provides the minimum and maximum scores for each of the models.

Table 1: Min and Max Scores of CAUS Block Scoring Models

Model	Min Score	Max Score
Decision Tree	0.59	29.33
Neural Network	-23.89	57.18
Regression	-123.57	294.30

Source: U.S. Census Bureau, Community Address Updating System Database, September 2010.

4.2. Linear Combination Assessment

As mentioned in Section 2, the main criteria used to assess potential linear combination was how well they ranked the top three to five percent of blocks, in terms of total adds, as compared to the top three to five percent of blocks as ranked by actual adds. The percent captured (PC) method was used to make this assessment. The following algorithm was used to implement this method for the study:

1) Define Optimal Outcome
    a. Compute the total number of adds in each block.
    b. Sort the blocks, from high to low, based on the number of adds.
    c. Compute the total number of adds in the top *X* percent of the sorted blocks. This represents the best possible outcome, in terms of total adds, that could be achieve when selecting *X* percent of the blocks in the data.

The data analysis for this paper was conducted using SAS software, Version 9.1.3 of the SAS System for Windows XP and Linux. Copyright © 2002-2003 SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

2) Compute Percent of Optimal Captured by Linear Combinations
    a. For each potential linear combination being assessed, sort the blocks, from high to low, based on the score assigned to them by that linear combination.
    b. Compute the total number of adds in the top *X* percent of the blocks as sorted by this score.
    c. Take the ratio of the value computed in step 2b to the value computed in step 1c and multiply it by 100. This gives the percent of the optimal total adds that the linear combination captured when it was used to select *X* percent of blocks in the data.

For this study, values of *X* equal to 3.0, 3.5, 4.0, 4.5, and 5.0 percent were considered.

The percent capture method was used because it allowed for comparisons between the different linear combinations while also providing information about how each linear combination compared to the best possible outcome, in terms of total adds, that could be achieve when selecting a certain percent of the blocks in the data.

Note that traditional regression model diagnostic tools were not used because the focus was on maximizing the total adds in the top 3-5 percent of block as scored by the different linear combinations, rather than choosing the model with the best overall fit.

5. Limitations

- Only linear combinations of model scores were considered due to the limited amount of time available to produce actionable results.
- This research included listing results the CAUS branch received from the field that had not yet been vetted by the GEO for inclusion into the MAF. Since these data do not necessarily reflect the true impact of the listings on the MAF, linear combinations based on them are potentially suboptimal in terms of MAF impact.

6. Results

6.1. Initial Stage

During this stage of the research, linear combinations were developed using both the raw and normalized model scores to see if normalization improved the performance of the linear combinations in terms of total adds in the top three to five percent of scored blocks. The coefficients computed using PROC REG, along with their associated t-distribution based p-values, are given in Table 2.

Table 2: Coefficients for Linear Combinations (Initial Stage)

	Raw		Normalized	
	Coefficient	**P-Value**	**Coefficient**	**P-Value**
Tree	0.22	<0.001	5.92	<0.001
Neural	0.10	<0.001	7.55	<0.001
Regression	-0.02	0.083	-6.54	0.084

Sources: U.S. Census Bureau, Community Address Updating System Database, September 2010; U.S. Census Bureau, Master Address File Updating Files, April 2010 – September 2010.

The percent captured method discussed in Section 4.2 was then used to compare these combinations to each other, to the original simple average based score, and to a random ordering of the data. The statistic compared was the percent of optimal adds captured in the top three to five percent of blocks as ordered by each method. Table 3 shows this comparison.

Each row in Table 3 gives the percent of total blocks being considered, the number of blocks being considered, the maximum number of adds possible in a selection of this size, and the percent of this optimal that is achieved using the various methods to select this number of blocks. An asterisk is used to identify the method(s) that achieve the highest percent captured for each percentile.

As an example of how to interpret Table 3, consider the first data row of the table. This row shows the results of a percent capture analysis based on approximately three percent of the 11,509 blocks in the initial data set. The statistic examined in this percent capture analysis is the number of adds per block.

- The value of 11,964 given in the **Adds in Optimal Selection** column is the total number of adds in the 345 blocks that had the most observed preliminary adds per block. This is the best possible selection of 345 blocks, in terms of total adds, obtainable from this data set.

- The value of 22.9 given in the **Original** column under the **Percent Capture Results** heading is the percent of the 11,964 optimal adds that the 345 blocks with the highest simple average based selection score contained.

- The value of 24.2 given in the **Raw Scores** column under the **Percent Capture Results** heading is the percent of the 11,964 optimal adds that the 345 blocks with the highest selection score as based on the linear combination developed using the non-normalized, or raw, model scores contained.

- The value of 23.9 given in the **Normed Scores** column under the **Percent Capture Results** heading is the percent of the 11,964 optimal adds that the 345 blocks with the highest selection score as based on the linear combination developed using the normalized model scores contained.

- The value of 8.5 given in the **Random Order** column under the **Percent Capture Results** heading is the percent of the 11,964 optimal adds that 345 blocks chosen at random contained. This column was included to allow a comparison between targeted and random block selection.

Table 3: Total Adds based Percent Capture Analysis (Initial Stage)

			Percent Capture Results			
Percent of Total	Number of Blocks	Adds in Optimal Selection	Original	Raw Scores	Normed Scores	Random Order
3.0	345	11,964	22.9	*24.2	23.9	8.5
3.5	402	12,910	25.5	*26.0	*26.0	8.9
4.0	460	13,798	26.3	26.9	*27.0	9.3
4.5	517	14,597	28.5	29.2	*29.7	9.5
5.0	575	15,336	29.9	31.4	*31.6	9.8

Sources: U.S. Census Bureau, Community Address Updating System Database, September 2010; U.S. Census Bureau, Master Address File Updating Files, April 2010 – September 2010.

The adds per unit and percent of blocks with adds were also examined for these blocks. The results are given in Tables 4 and 5. These tables do not reflect a percent captured analysis based on adds per unit and percent of blocks with adds, but give these statistics for the blocks selected as part of the total adds based percent captured analysis shown in Table 3.

As an example of how to interpret Table 4, consider the first data row of the table.

- The value of 0.49 given in the **Blocks with the Most Adds (from Table 3)** column is the number of adds per preexisting unit in the 345 blocks that had the most observed preliminary adds per block. Note that this value does not represent the best possible adds per unit that could be achieved by selecting 345 blocks, but the adds per unit of the 345 blocks from the first data row of Table 3 that had 11,964 adds.

- The value of 23.8 given in the **Original** column under the **Percent Capture Results** heading is the adds per preexisting unit, expressed as a percent of 0.49, that the 345 blocks with the highest simple average based selection score contained.

- The value of 23.4 given in the **Raw Scores** column under the **Percent Capture Results** heading is the adds per preexisting unit, expressed as a percent of 0.49, that the 345 blocks that had the highest selection score as based on the linear combination developed using the non-normalized, or raw, model scores contained.

- The value of 23.2 given in the **Normed Scores** column under the **Percent Capture Results** heading is the adds per preexisting unit, expressed as a percent of 0.49, that the 345 blocks that had the highest selection score as based on the linear combination developed using the normalized model scores contained.

- The value of 18.6 given in the **Random Order** column under the **Percent Capture Results** heading is the adds per preexisting unit, expressed as a percent of 0.49, that the 345 blocks randomly chosen for Table 3 contained.

Table 4: Adds per Preexisting Unit Comparison (Initial Stage)

			Percent Capture Results			
Percent of Total	**Number of Blocks**	**Blocks with the Most Adds (from Table 3)**	**Original**	**Raw Scores**	**Normed Scores**	**Random Order**
3.0	345	0.49	*23.8	23.4	23.2	18.6
3.5	402	0.46	*26.1	25.0	25.0	19.3
4.0	460	0.43	*26.2	25.6	25.6	20.4
4.5	517	0.41	*27.8	27.4	*27.8	20.5
5.0	575	0.38	29.4	29.6	*29.8	21.5

Sources: U.S. Census Bureau, Community Address Updating System Database, September 2010; U.S. Census Bureau, Master Address File Updating Files, April 2010 – September 2010.

As an example of how to interpret Table 5, consider the first data row of the table.

- The value of 100.0 given in the **Blocks with the Most Adds (from Table 3)** column is the percent of blocks with adds in the 345 blocks that had the most observed preliminary adds per block.

- The value of 78.8 given in the **Original** column under the **Percent Capture Results** heading is the percent of blocks with adds in the 345 blocks with the highest simple average based selection score relative to the 100.0 percent with adds in the **Blocks with the Most Adds (from Table 3)** column.

- The value of 82.6 given in the **Raw Scores** column under the **Percent Capture Results** heading is the percent of blocks with adds in the 345 blocks that had the highest selection score as based on the linear combination developed using the non-normalized, or raw, model scores relative to the 100.0 percent with adds in the **Blocks with the Most Adds (from Table 3)** column.

- The value of 82.0 given in the **Normed Scores** column under the **Percent Capture Results** heading is the percent of blocks with adds in the 345 blocks that had the highest selection score as based on the linear combination developed using the normalized model scores relative to the 100.0 percent of blocks with adds in the **Blocks with the Most Adds (from Table 3)** column.

- The value of 53.6 given in the **Random Order** column under the **Percent Capture Results** heading is the percent of blocks with adds in the 345 blocks randomly chosen for Table 3 relative to the 100.0 percent of blocks with adds in the **Blocks with the Most Adds (from Table 3)** column.

Table 5: Percent of Blocks with Adds Comparison (Initial Stage)

			Percent Capture Results			
Percent of Total	Number of Blocks	Blocks with the Most Adds (from Table 3)	Original	Raw Scores	Normed Scores	Random Order
3.0	100.0	0.49	78.8	*82.6	82.0	53.6
3.5	100.0	0.46	77.1	80.8	*81.6	54.0
4.0	100.0	0.43	76.5	80.0	*80.2	52.8
4.5	100.0	0.41	77.8	80.3	*80.5	53.2
5.0	100.0	0.38	77.9	80.3	*80.5	53.6

Sources: U.S. Census Bureau, Community Address Updating System Database, September 2010; U.S. Census Bureau, Master Address File Updating Files, April 2010 – September 2010.

Tables 3, 4, and 5 suggest that using an updated linear combination based on the either the raw or normed scores to select blocks will increase the number of adds obtained by listing. In addition, this should result in selecting more blocks with adds. However, they also suggest that the number of adds per preexisting unit will decrease. Of the two updated linear combination, the one based on normed model scores generally outperformed the one based on raw model scores. Note that all selection scores outperformed a random ordering of blocks for all statistics.

This analysis led to the selection of the linear combination below for the initial stage:

$$S_i = 5.92 \cdot ND_i + 7.55 \cdot NN_i - 6.54 \cdot NR_i, \qquad (6.1)$$

where

ND_i = normalized decision tree based score for block i,
NN_i = normalized neural network based score for block i,
NR_i = normalized regression based score for block i.

The resulting block score was used to select the blocks that were listed through the CAUS program during the March 2011 – August 2011 listing period.

6.2. Final Stage

During this stage of the research we compared linear combinations developed using PROC REG based only on normalized model scores. The coefficients computed, along with their associated p-values, are given in Table 6.

Table 6: Coefficients for Linear Combination (Final Stage)

	Coefficient	P-Value
Tree	4.99	<0.001
Neural	4.12	<0.001
Regression	-3.03	0.121

Sources: U.S. Census Bureau, Community Address Updating System Database, September 2010; U.S. Census Bureau, Master Address File Updating Files, September 2010 – January 2011; U.S. Census Bureau, Master Address File Transaction File, January 2011.

One thing that can be seen from looking at Tables 2 and 6 is that the coefficients for the regression score were only marginally significant in both phases of this research. Given that data were from a non-random sample and previous work done by the CAUS branch suggested that adds per block follows a non-normal distribution with heteroscedastic variance (Lawrence, 2010), it would be questionable to use the p-value given by PROC REG as a decision criterion. However, this did suggest that the regression scoring model, given the information from the decision tree and neural network based models, may not significantly improve the CAUS branch's ability to predict the number of adds in a block.

To examine this further, another linear combination was developed in PROC REG that only used the normalized scores of the decision tree and neural network based models as inputs. The coefficients computed, along with their associated p-values, are given in Table 7.

Table 7: Coefficients of Linear Combination without Regression (Final Stage)

	Coefficient	P-Value
Tree	5.27	<0.001
Neural	1.45	<0.004

Sources: U.S. Census Bureau, Community Address Updating System Database, September 2010; U.S. Census Bureau, Master Address File Updating Files, September 2010 – January 2011; U.S. Census Bureau, Master Address File Transaction File, January 2011.

The percent captured method discussed in Section 4.2 was then used to compare these combinations to each other, to the results of the initial rescoring, to the original equal factor score, and to a random ordering of the data. The statistic compared was the percent of optimal adds captured in the top three to five percent of blocks ordered by each method. Table 8 shows this comparison.

Each row in Table 8 gives the percent of total blocks being considered, the number of blocks being considered, the maximum number of adds possible in a selection of this size, and the percent of this optimal that is achieved using the various methods to select this number of blocks. An asterisk is used to identify the method(s) that achieve the highest percent captured for each percentile.

The interpretation of the values in Table 8 is the same as that for Table 3 in Section 6.1.

Table 8: Total Adds Bases Percent Captured Analysis (Final Stage)

			Percent Capture Results				
Percent of Total	Number of Blocks	Adds in Optimal Selection	Original	Initial Refresh	Final Refresh (Full)	Final Refresh (no Reg.)	Random Order
3.0	358	10,133	20.4	*20.7	*20.7	20.4	5.4
3.5	418	10,885	22.4	23.0	*23.2	23.1	5.9
4.0	478	11,558	23.3	24.5	*25.0	24.6	6.2
4.5	538	12,159	25.8	*27.7	*27.7	27.4	6.7
5.0	598	12,721	27.5	*29.0	28.9	*29.0	7.4

Sources: U.S. Census Bureau, Community Address Updating System Database, September 2010; U.S. Census Bureau, Master Address File Updating Files, September 2010 – January 2011; U.S. Census Bureau, Master Address File Transaction File, January 2011.

The adds per unit and percent of blocks with adds were also examined for these blocks. The results are given in Tables 9 and 10. As with the initial stage of this research, these tables do not reflect a percent captured analysis based on adds per unit and percent of blocks with adds, but give these statistics for the blocks selected in the total adds percent captured analysis shown in Table 8. Thus 0.39 is not the best adds per unit you could achieve by selecting 358 blocks, but the adds per unit of the 358 blocks from Table 8 that had 10,133 adds.

The interpretation of the values in Table 9 is the same as that for Table 4 in Section 6.1.

Table 9: Adds per Preexisting Unit Comparison (Final Stage)

			Percent Capture Results				
Percent of Total	Number of Blocks	Blocks with the Most Adds (from Table 8)	Original	Initial Refresh	Final Refresh (Full)	Final Refresh (no Reg.)	Random Order
3.0	358	0.39	*21.9	20.7	20.7	20.5	12.4
3.5	418	0.36	*23.8	22.8	22.9	23.4	13.7
4.0	478	0.34	23.8	23.7	24.2	*24.3	14.1
4.5	538	0.31	26.4	26.9	27.0	*27.1	15.4
5.0	598	0.30	28.0	28.0	28.0	*28.4	16.4

Sources: U.S. Census Bureau, Community Address Updating System Database, September 2010; U.S. Census Bureau, Master Address File Updating Files, September 2010 – January 2011; U.S. Census Bureau, Master Address File Transaction File, January 2011.

The interpretation of the values in Table 10 is the same as that for Table 5 in Section 6.1.

Table 10: Percent of Blocks with Adds Comparison (Final Stage)

Percent of Total	Number of Blocks	Blocks with the Most Adds (from Table 8)	Percent Capture Results: Original	Initial Refresh	Final Refresh (Full)	Final Refresh (no Reg.)	Random Order
3.0	358	100.0	68.4	*69.3	*69.3	68.7	29.6
3.5	418	100.0	66.3	*68.2	*68.2	67.9	29.7
4.0	478	100.0	65.9	67.4	*67.6	67.4	30.5
4.5	538	100.0	66.9	67.8	67.8	*68.0	30.5
5.0	598	100.0	66.1	*68.2	68.1	67.7	31.3

Sources: U.S. Census Bureau, Community Address Updating System Database, September 2010; U.S. Census Bureau, Master Address File Updating Files, September 2010 – January 2011; U.S. Census Bureau, Master Address File Transaction File, January 2011.

Tables 8, 9, and 10 suggest that the linear combination of all the normalized model scores developed using the final adds data is a marginal improvement over the linear combination developed using the preliminary adds data. They also suggest that a linear combination with the regression model score included performs slightly better, with the exception of adds per unit. Thus we would expect the linear combination with regression included to result in a higher number of total adds, but lower adds per unit. From the above, we did not feel there was enough evidence to justify removing the regression model from the linear combination.

This analysis let to the selection of the linear combination below:

$$S_i = 4.99 \cdot ND_i + 4.12 \cdot NN_i - 3.03 \cdot NR_i, \tag{6.2}$$

where

ND_i = normalized decision tree based score for block i,
NN_i = normalized neural network based score for block i,
NR_i = normalized regression based score for block i.

The resulting block score was used to select the blocks that were listed through the CAUS program during the September 2011 – February 2012 listing period.

7. Future Research

As part of its ongoing block targeting research the CAUS branch should look into which of these three models, or some combination thereof, should be further developed for use in future block targeting. The fact that the coefficient for the regression model was only marginally significant in the final linear combination provides some evidence that including information from a regression based model may not be necessary when selecting blocks for listings.

8. References

Bates, Lawrence. (2011). “Editing the MAF Extracts and Creating the Unit Frame Universe for the American Community Survey (2012 Main Phase).” Internally distributed software specification published as *DSSD 2012 American Community Survey Memorandum Series #ACS12-UC-1*, June 2, 2011. Washington, DC: U.S. Census Bureau.

Bishop, Christopher M. (1995). *Neural Networks for Pattern Recognition*. Oxford, U.K.: Oxford University Press.

Gabriel, K. Ruben. (1978). “A Simple Method of Multiple Comparisons of Means.” *Journal of the American Statistical Association*. Vol.73, No. 364, December 1978, pp. 724-729.

GEO (Geography Division). (2011a). “Address Characteristic Type Software Requirement Specification.” Internally distributed software specification, version 1.5, last modified February 3, 2011. Washington, DC: U.S. Census Bureau.

———. (2011b). “Delivery Sequence File Refresh Software Requirements Specification.” Internally distributed software specification, version 1.8, last modified July 28, 2011. Washington, DC: U.S. Census Bureau.

Kass, G.V. (1980). “An Exploratory Technique for Investigating Large Quantities of Categorical Data.” *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, Vol. 29, No. 2, 1980 pp. 119-127.

Lawrence, James and Holly Malanowski. (2010) “CAUS Branch, Q1 2010 Block Targeting Documentation (DRAFT Version 7).” Unreleased memorandum, last modified July 02, 2010. Washington D.C.: U.S. Census Bureau.

Larose, Daniel T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*. Hoboken, NJ: John Wiley & Sons, Inc.

Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar (2005). *Introduction to Data Mining*. Boston, MA: Pearson Education, Inc.

SAS Institute Inc. (2003). “Enterprise Miner 4.3 Reference Help.” Reference document accessible through the SAS Enterprise Miner 4.3 software. Cary, NC: SAS Institute Inc.

Attachment A

Brief Summary of Modeling Methods used to Score Blocks for Community Address Updating System Listings since September 2010

1. Introduction

The CAUS branch currently uses information from three block scoring models it developed using SAS® Enterprise Miner™ 4.3 software to select blocks for listing through the CAUS program. The models are based on the following methods:

- Decision Tree Learning
- Neural Network Analysis
- Regression Analysis

A brief summary of the data and methodology used to develop these models is provided below. Note that much of the methodology information in this summary comes from the Enterprise Miner 4.3 Reference Help documentation bundled with the software. Additional information about these methods can be found in Bishop (1995), Larose (2005), and Tan, Steinbach, and Kumar (2006).

2. Data Used

The parameters in these models were estimated by using 2007 and 2008 block and county characteristic data to predict the number of adds observed in blocks during the 2010 Decennial Address Canvassing operation conducted in the spring and summer of 2009. A training and a validation data set were used in model development. These data sets were created by randomly assigning 70 percent of blocks to the training data set, and the remaining 30 percent of the blocks to the validation data set. The training data set was used to estimate the initial parameters of a potential block scoring model. The validation data set was then used to adjust the resulting model as necessary. This two-step process helped ensure that the resulting block scoring models generalized well, and were not overfitting to the peculiarities of a particular data set.

The following block level information from the July 2007 and July 2008 vintages of the MAF was used in model development:

- ACT code of blocks
- The number and type of existing MAF housing units in blocks
- The growth in housing units in blocks between 2007 and 2008
- Whether or not blocks were classified as urban or rural
- The physical area of blocks
- The housing unit density of blocks

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

- Whether addresses in blocks had undergone non-city-style to city-style conversions under the USPS's Locatable Address Conversion System (LACS)
- How blocks were enumerated in the 2000 Decennial

The following information from spring 2007 and spring 2008 vintages of the DSF was used at both the block and county level when developing the models:

- The number and growth of all DSF delivery points
- The number and growth of city-style delivery points
- The number and growth of rural-route delivery points
- The number and growth of PO Box delivery points
- The number of addresses that had undergone LACS conversions

In addition, the difference between the number and growth of MAF and DSF units was used in model development.

Updated versions of these block and county characteristics are used when blocks are periodically scored for selection.

3. Decision Tree Learning

The decision tree learning method that was used to develop this scoring model tries to find the best way to use block characteristic information to partition the blocks in a data set. It does this by analyzing the difference in the average of the target variable, adds-per-block, of possible partitions. Note that the default decision tree learning algorithm of the SAS Enterprise Miner 4.3 software was used to develop this scoring model. The below provides a brief summary of the default decision tree learning algorithm of the SAS Enterprise Miner 4.3 software. A more detailed description of the algorithm can be found in the reference documentation included with the software (SAS Institute Inc., 2003).

The first part of the decision tree learning algorithm is called the split search. The split search starts by selecting a characteristic for partitioning the available training data. If the measurement scale of the selected characteristic is interval, each unique value serves as a potential split point for the data. If the characteristic is categorical, the average value of the target is taken within each categorical input level. The averages serve the same role as the unique interval characteristic values in the discussion that follows.

For each potential split point of the selected block characteristic, two groups are generated. Cases with characteristic values less than the potential split point are said to branch left. Cases with characteristic values greater than the potential split point are said to branch right. The average of the target variable is computed in each branch, and an *F* statistic is computed based on the differences of these averages. Large values for the *F* statistic suggest that the average value in the left branch is different from the average value in the right branch. A large difference in these averages indicates a good split.

The *p*-value for this *F* statistic is then computed. The *p*-value indicates the likelihood of obtaining this *F* value under the hypothesis that the means in both branches are equal. This *p*-value is also based on the assumption that the underlying data are normally distributed with equal variance. The *p*-value of each split is then adjusted by the SAS Enterprise Miner 4.3 software to account for the multiple comparisons conducted during the split search process. The software uses KASS (Kass, 1980) or Gabriel (Gabriel, 1978) adjustments and Bonferroni adjustments when doing this (SAS Institute Inc., 2003). For larger data sets, these adjusted *p*-values can be very close to zero. For this reason, the quality of a split is reported by *logworth* = -*log*(adjusted *F*-test *p*-value).

The logworth of at least one potential split of a block characteristic must exceed a threshold for a split to occur with that characteristic. For our model, this threshold corresponds to an adjusted *F*-test *p*-value of 0.20 or a logworth of approximately 0.70. That is, for a potential split to be considered significant there must be a less than a 0.20 adjusted probability that the observed difference in the means would occur by chance under the hypothesis that they are equal.

After determining the best split for every block characteristic, the tree algorithm compares each best split's corresponding logworth. The best split with the highest logworth is selected. The training data is then partitioned along that best split into two branches, and the split search process is repeated on each of the resulting branches. The entire process is repeated until no split of a block characteristic with an adjusted logworth value above the threshold value can be found within any branch of the tree. The resulting tree is called the maximal tree.

Note that for our model, we only considered decision trees with a maximum number of two branches at each split, and a maximum depth of six splits from the full data set to a final branch. In addition, a branch of the tree had to contain at least one percent of the total data for a split search to be conducted within it, and any split could not result in a branch that contained less than 0.1 percent of the total data.

The tree was then pruned using the validation data set. During this process, all possible subtrees of the maximal tree are created on the validation data set. These subtrees are created by working backwards from the maximal tree; that is, each split of the maximal tree is collapsed starting from the lowest level splits to the highest. For each of these subtrees, the average squared error between its predicted adds-per-block values and the observed adds-per-block values is computed. The tree with the average lowest squared error on the *validation* data set is selected as the final decision tree. Note that any time two potential trees have the same squared error, the simplest tree, in terms of the fewest number of splits, is preferred.

Only splits based on the total number of 2008 MAF housing units in a block were included in this model after the pruning process.

4. Neural Network Analysis

The neural network we developed was a multilayer perceptron (MLP). In general, a MLP can be thought of as an indirect regression model. The target variable, adds-per-block, was fitted to a linear regression model that used functions of the original block characteristic variables as inputs. These intermediate functions are traditionally referred to as hidden units. We used hyperbolic tangent functions for the hidden units of our neural network. The hyperbolic tangent function is commonly used for MLPs, and is the default option in the SAS Enterprise Miner 4.3 software. More about the theory and applications of MLP neural networks can be found in Bishop (1995).

Symbolically, this type of neural network can be expressed as:

$$y_i = w_{00} + w_{01} \cdot H_{i1} + w_{02} \cdot H_{i2} + \cdots + w_{0n} \cdot H_{in},$$

where

$$\begin{aligned} &H_{i1} = \tanh\ w_{10} + w_{11}x_{i1} + w_{12}x_{i2} + \cdots + w_{1m}x_{im}\ \ , \\ &H_{i2} = \tanh\ w_{20} + w_{21}x_{i1} + w_{22}x_{i2} + \cdots + w_{2m}x_{im}\ \ , \\ &\vdots \\ &H_{in} = \tanh\ w_{n0} + w_{n1}x_{i1} + w_{n2}x_{i2} + \cdots + w_{nm}x_{im}\ \ . \end{aligned}$$

The $x_{i1}, x_{i2}, \dots, x_{im}$ are the original characteristic variables for block i, $H_{i1}, H_{i2}, \dots, H_{in}$ are the hidden units for block i, and y_i is the predicted number of adds for block i. The w's are the parameters being estimated by the neural network.

Our neural network was developed using the default options in the SAS Enterprise Miner 4.3 software, and was initialized in the following way:

- All interval block characteristic variables, the x_{ij}'s, were standardized by subtracting the mean value of the input variable from all observations and dividing by their standard deviation. Thus all input variables have a mean of zero and a standard deviation of one.
- Only three hidden units, $H_1, H_2,$ and H_3 were considered.
- All input connection weights, those applied to the x_{ij}'s, and input biases, $w_{10}, w_{20}, \dots, w_{n0}$, were initialized by drawing a pseudo-random value from a N(0, 1) distribution.
- All output connection weights, $w_{01}, w_{02}, \dots, w_{0n}$, were initialized at zero.
- w_{00} was initialized as tanh^{-1} y .

The y_i's are then estimated on the training data set using these initial values, and the sum of squared errors (SSE) and mean squared errors (MSE) are computed for these initial estimates. A vector of partial derivates of the SSE with respect to each of the w's is computed, and this information is used to adjust the w's in order to reduce the SSE of the estimates. These new w's are then used to produce another set of y_i's, for which the SSEs and MSEs are computed. The optimization process is repeated until either a predetermined amount of time elapses, the maximum number of iterations is reached, or convergence criteria are met.

In order to prevent overfitting to the data in the training data set, the validation data set is used to select the final model. Each time the algorithm computes a new set of w's based on the training data set, it uses those w's to estimate y_i's based on validation data set inputs. The adds-per-block information from the validation data set is then used to compute the MSEs for these y_i's. Of the models created using the data in the training data set, the one with the lowest MSE on the *validation* data set is selected as the final model. Note that all block characteristic variables were included in the final model.

5. Regression Analysis

The linear regression model developed by the CAUS branch for block targeting was based on the traditional assumptions of regression analysis:

- The relationship between the dependent and independent variables is linear.
- Observation errors are independent.
- Observation errors are homoscedastic.

An identify link function was used between the dependent and explanatory variables. All available block characteristic variables were included in this model, but no interaction terms were considered. The parameters chosen were those that minimized the square error of the model on the *training* data set.

Attachment B

Community Address Updating System Universe

When selecting blocks for listing, the Community Address Updating System (CAUS) branch first considers the address characteristic type (ACT) code assigned to blocks by the Geography Division (GEO). This code indicates both the type of addresses in a block and how many of a block's addresses can be matched to addresses on the U.S. Postal Service's Delivery Sequence File (DSF) (GEO, 2011a). Note that a block's ACT code may change from year to year based on changes in its MAF information. The CAUS branch classifies blocks into one of three groups based on their ACT code.

- Yes, or "Y", blocks are considered to be in the CAUS universe and eligible for listing.
- Maybe, or "M", blocks are considered to be in the margins of the CAUS universe.
- No, or "N", blocks are considered to be outside the CAUS universe.

Blocks in the "N" group contain either only city-style addresses, non-residential addresses, or no addresses. Blocks in the "M" group contain a mixture of city-style and non-city-style addresses where some of the addresses match to a DSF delivery point and at least 80 percent of the addresses are city-style. Blocks in the "Y" group contain either only non-city-style addresses, a mixture of city-style and non-city-style addresses where either none or all of the addresses match to a DSF delivery point, or a mixture of city-style and non-city-style addresses where some of the addresses match to a DSF delivery point and less than 80 percent of the addresses are city-style. Table B-1 gives the ACT codes that belong to each group.

Table B-1: ACT Code Groups Used for CAUS Listings

CAUS Group	ACT Codes
Y	M1, ME, MF, MG, M3, N1, N2, N3, P1, P2, P3, R1, R2, R3
M	MA, MB, MC, MD
N	B1, B2, B3, C1, C2, C3, Z0

Attachment C

Address Characteristic Type Code Definitions

The Address Characteristic Type (ACT) code is a two-character code that describes the type of Master Address File (MAF) addresses in the block, and indicates how many of the block's MAF addresses can be matched to delivery points on the US Postal Service's Delivery Sequence File (DSF) (GEO, 2011a).

Table C-1: ACT Code Definitions with CAUS Groups

CAUS Group	ACT Code	Definition
Y	D1	Description only, MAF description only addresses cannot be matched to DSF addresses
	M1	City-style and noncity-style, no addresses matched to DSF
	ME	City-style and noncity-style, some addresses matched to DSF where the percent City-style is in [75, 80)
	MF	City-style and noncity-style, some addresses matched to DSF where the percent City-style is in [70, 75)
	MG	City-style and noncity-style, some addresses matched to DSF where the percent City-style is in (0, 70)
	M3	City-style and noncity-style, all addresses matched to DSF
	N1	Assorted noncity-style, no addresses matched to DSF
	N2	Assorted noncity-style, some addresses matched to DSF
	N3	Assorted noncity-style, all addresses matched to DSF
	P1	PO Box, no addresses matched to DSF
	P2	PO Box, some addresses matched to DSF
	P3	PO Box, all addresses matched to DSF
	R1	Rural Route, no addresses matched to DSF
	R2	Rural Route, some addresses matched to DSF
	R3	Rural Route, all addresses matched to DSF
M	MA	City-style and noncity-style, some addresses matched to DSF where the percent City-style is in [95, 100)
	MB	City-style and noncity-style, some addresses matched to DSF where the percent City-style is in [90, 95)
	MC	City-style and noncity-style, some addresses matched to DSF where the percent City-style is in [85, 90)
	MD	City-style and noncity-style, some addresses matched to DSF where the percent City-style is in [80, 85)
N	B1	Non-residential only, no addresses match to DSF
	B2	Non-residential only, some addresses matched to DSF
	B3	Non-residential only, all addresses matched to DSF
	C1	City-style, no addresses match to DSF
	C2	City-style, some addresses match to DSF
	C3	City-style, all addresses match to DSF
	Z0	No addresses