Variable selection for multinomial logistic regression modeling to assign one of six census mindsets using Big Data

Mary H. Mulry, Yazmin A. Garcia Trejo, and Nancy Bates[1]
U.S. Census Bureau, Washington, DC 20233

Abstract
The U.S. Census Bureau is preparing to field the 2020 Census Communications Campaign to encourage participation in the 2020 Census. Similar campaigns aided in maintaining high self-response rates for the 2000 and 2010 Censuses. To prepare, the U.S. Census Bureau fielded the 2020 Census Barriers, Attitudes and Motivators Study (CBAMS) sample survey to collect data on attitudes and knowledge about the U.S. Census. Data from over 17,000 respondents was used to classify individuals into one of six psychographic profiles referred to as Census "mindsets". In social marketing campaigns, mindsets are constructed to reflect an individual's knowledge, attitudes and opinions toward a topic. The mindsets are then used in developing messages with a call to action. In our case, the requested action is a response to the 2020 Census. Our research examines the feasibility of assigning a mindset to each record in a Big Data file, which is a third-party dataset containing over 250 million adult records and ultimately to households. The 2020 CBAMS variables used in determining the mindsets are not present on the third-party dataset although the dataset does contain over 500 variables that reflect demographics, socioeconomic status, attitudes and behavior. Our approach links the 2020 CBAMS survey records to the third-party dataset and then uses multinomial logistic regression with independent variables from the third-party dataset to predict the probabilities of the mindsets.

Key Words: 2020 Census, 2020 Census Communications Campaign, self-response, 2020 Census Barriers, Attitudes and Motivators Study

1. Background

The Census communications campaign is a massive endeavor organized every ten years. The ultimate goal of the communications campaigns is to persuade, every person in the U.S. to participate in the Census. As part of this endeavor, the Census researchers conducted a foundational research project called the 2020 Census Barriers, Attitudes and Motivators Study (a nationally representative public opinion sample survey and focus groups) to inform through empirical evidence the advertising campaign and the classification of the U.S. population based on their probability to participate in the 2020

[1] This report is released to inform interested parties and encourage discussion of work in progress. The views expressed on statistical, methodological, and operational issues are those of the authors and not those of the U.S. Census Bureau. Data approved by CBDRB-FY19-451. July 15, 2019.

Census. This paper is part of this foundational research effort and it aims on using the public opinion data (2020 CBAMS survey) to document efforts to implement a typing tool methodology that matches individual and household level data and the characteristics that will make them more (or less) likely to participate in the Census.

Our research builds on experience of past census communications programs to explore the feasibility of assigning a mindset to each record in a Big Data file, which is a third-party dataset, called the National Household File (NHF), containing over 250 million adult records and ultimately to households. This means that we match CBAMS survey individual respondents to a single individual from a household identified in the third-party dataset. We were able to connect individuals from the CBAMS sample survey with the third-party dataset because both datasets had address and demographics. The 2020 CBAMS variables used in determining the mindsets are not present on the third-party dataset although the dataset does contain over 500 variables that reflect demographics, socioeconomic status, attitudes and behavior. Our approach links the 2020 CBAMS survey records to the third-party dataset and then uses multinomial logistic regression with independent variables from the third-party dataset to predict the probabilities of the mindsets.

Our research investigates which variables from the third-party dataset are important in predicting mindset assignment. In addition, we explore fitting a multinomial logistic regression model to predict the mindsets using variables from the third-party dataset. We examine the methodology for using estimated probabilities for the mindsets predicted by the multinomial logistic regression model to make a household-level assignment of psychographic “mindsets,” which was one of the approaches originally suggested but not selected for use in the 2020 Census Communication Campaign. Section 2 provides background information related to the overview of the communications campaign, and Section 3 discusses the planned methodology for using the mindsets in the communications campaign, which we investigate in this study, and the alternative methodology developed for use in the 2020 Census campaign. Section 4 discusses the preparation of the data used in our research while Section 5 contains a description of the development of the ‘Typing Tool’ and the variables important for distinguishing the mindsets that we subsequently use as candidate variables in the model fitting. The discussion of the variable selection and the multinomial logistic regression model fit with the selected variables appears in Section 6. Section 7 contains a discussion of using the estimated probabilities of the mindsets for the residents of a household to assign a mindset to the household. Section 8 describes our conclusions.

2. Overview of 2020 Communications Campaign

The goal of the 2020 Census Integrated Communications Campaign (ICC) is to encourage self-response in the 2020 Census through a research-based communications campaign. The Census Bureau engaged in similar social marketing campaigns that included a paid advertising campaign in both the 2000 and 2010 Census. Both censuses achieved a mail response rate of 67 percent although the budgeted mail response rate was 61 percent for the 2000 Census and 64 percent for the 2010 Census (Bates 2017). Experts believe the paid campaigns played a large role in this success in both censuses (Bates, 2017; Evans, Douglas, Datta, and Yan 2014; Williams, Bates, Lotti, and Wroblewski 2014). Such campaigns include paid advertising (television, radio, print, digital, etc.) as well as earned media (such as newspaper articles and news segments) and a more community-based outreach using Partnership Specialists who partner with local elected officials, community activists, leaders and advocates to raise awareness of the Census and encourage participation.

Because the 2020 Census must count every single person living in the U.S. on April 1st, the communications and advertising campaign must be extremely robust to reach all segments of such a diverse population. To help with creative message development, social marketing campaigns commonly develop psychographic profiles of the population (known as "mindsets") according to their knowledge, attitudes, and practices towards a particular product (or in our case the 2020 Census). The 2020 Census advertising contractor, Young and Rubicam (Y&R), used results from a nationwide sample survey (the 2020 Census Barriers, Attitudes, and Motivators Study or CBAMS) to produce six such mindsets that reflect shared patterns of attitudes, behaviors, and motivators toward the 2020 Census (See Kulzick et al 2019).

The Census Bureau administered the 2020 CBAMS survey between February 20, 2018 and April 17, 2018 to 50,000 housing units in all 50 states and the District of Columbia. The survey contained questions designed to measure the public's attitudes, knowledge, and opinions regarding the 2020 Census. The results were primarily for the purposes of developing the creative platform and messaging for the 2020 Census Communications Campaign.

The sample design for the survey included stratifying the U.S. population into eight strata based on a Census tract's racial and ethnic makeup as well as characteristics related to internet response in the American Community Survey. Each household in the sample received a prepaid incentive and up to five mailings inviting them to participate by mail or Internet in either English or Spanish (for more information on this methodology, see McGeeney et al., 2019). In all, approximately 17,500 adults responded to the survey, and survey weights were constructed so that the weighted distribution of the respondents matched the distribution of all householder adults in the U.S. The final, weighted response rate was 39.4 percent and was calculated using a modified version of the American Association for Public Opinion Research (AAPOR) RR3 (AAPOR 2016).

2.1 Methodology of the Mindsets

Since the number of potential inputs to mindset segmentation was large (45+ survey items), a data reduction technique was necessary[2]. This included three steps:

1) **Dimension Reduction** – The knowledge, attitudes, barriers, and motivators measured in 2020 CBAMS sample survey reflected a smaller number of underlying factors. The team used principal component analysis (PCA) to reduce 2020 CBAMS survey variables to a smaller number of factors that captured most of the information in the responses. The PCA suggested that the optimal number of principal components was eight, as indicated by a scree plot of eigenvalues. Next, varimax rotation was used to ensure that each variable corresponded to a single, uncorrelated factor. Finally, mean factor loadings across the eight factors were calculated for each case.

2) **Candidate Identification** – The team then created candidate mindset solutions using a clustering algorithm to group respondents based on underlying similarities in the

[2] This is a summary about the modeling process to determine the mindset segmentation. Specific details about the methodology will be available in a forthcoming report entitled "2020 Census Predictive Models and Audience Segmentation Report (Kulzick et al 2019).

factors established in the previous step. The Ward's clustering identified a preset number of mindsets, so the team developed separate solutions for different numbers of mindsets. We evaluated candidate solutions using observations and metrics such as Dunn's index (Xiong and Li 2013), which measures the compactness and separation of the clusters in a solution. This process identified three sets of solutions while considering the schedule and available budget for analysis.

3) **Final Selection** – The final selection process determined the mindsets that are actionable for the communications team. An evaluation team composed of Census staff, Team Y&R strategists, media planners, and creatives—including those from the multicultural agency partners—reviewed three candidate mindset solutions and selected the most actionable solution for the communications campaign. Segments were considered actionable if their identification guides decision-making for the effective specification of marketing instruments (Wedel and Kamakura 2012). The review of these candidate solutions involved the evaluation of the 2020 CBAMS survey questions with the most distinctive set of responses for each potential mindset. The process identified six distinct mindsets:

Eager Engagers are the most civically engaged mindset and have the highest knowledge about the census, as well as intent to respond. This mindset also comprises the highest percentage of college-educated people and the highest household incomes.

Fence Sitters are the largest mindset in number of CBAMS respondents. They do not have major concerns about taking the census and are less civically active than Eager Engagers, but they are still highly inclined to respond. This mindset is the least diverse and has the highest percentage of males.

Confidentiality Minded are most concerned that their answers to the census will be used against them, but they believe their answers matter and are still fairly likely to respond. This mindset is the most diverse and has the highest percentage of foreign-born people.

Head Nodders are most likely to give affirmative answers to all knowledge questions and demonstrate significant knowledge gaps in specific areas. This mindset has the highest percentage of people 18-34 years old and above average percentage of foreign-born people.

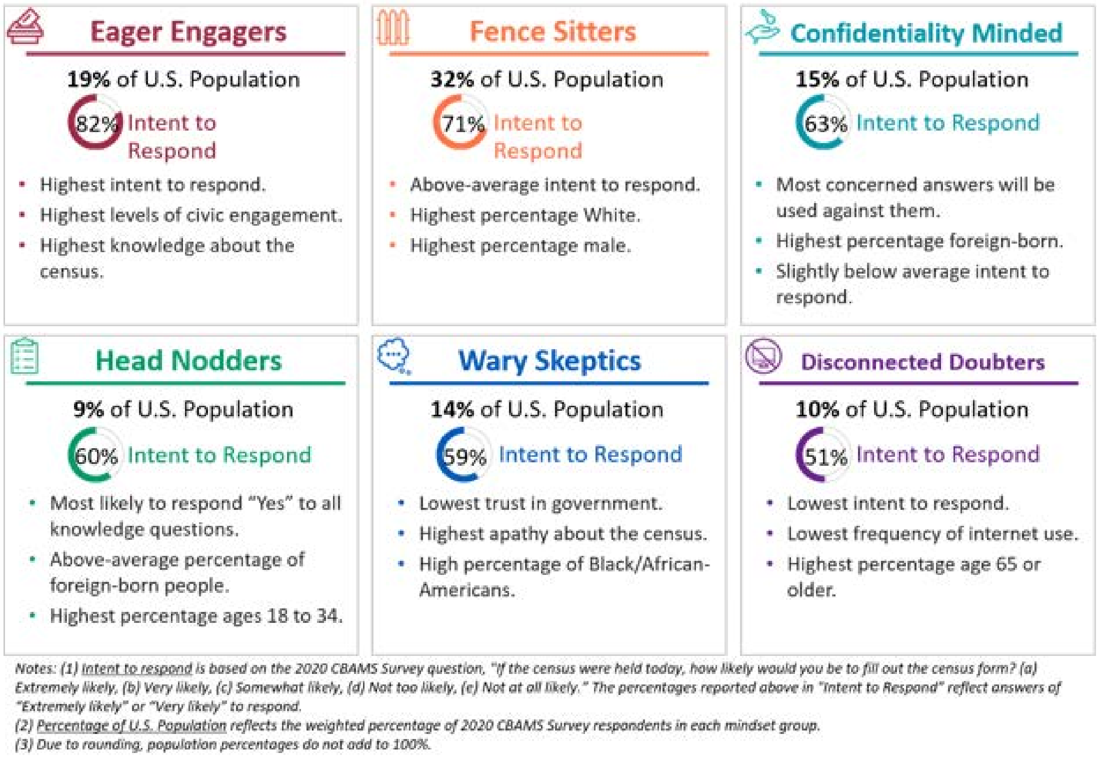
Wary Skeptics are skeptical of the government, as shown by their high distrust of the government, and are, therefore, reluctant to participate in the census. This mindset has the highest percentage of Black/African-Americans and below average education attainment.

Disconnected Doubters do not use or have access to the internet, do not believe that their response matters, and are the least likely to respond to the census. This mindset has the highest percentage of people 65 years or older, and has the lowest levels of education.

The six mindsets were structured around the following four questions:

- **Who are they?**
- **Do they intend to respond, and how do they think about the census?**
- **What are their potential barriers to participation?**
- **What are their potential motivators for participation?**

Figure 1 is a visual dashboard of key characteristics of the final six mindsets, which, in order of the percentage who intend to respond, are: (1) Eager Engagers, (2) Fence Sitters, (3) Confidentiality Minded, (4) Head Nodders, (5) Wary Skeptics, and (6) Disconnected Doubters (McGeeney et al 2019).



Eager Engagers

19% of U.S. Population

82% Intent to Respond

- Highest intent to respond.
- Highest levels of civic engagement.
- Highest knowledge about the census.

Fence Sitters

32% of U.S. Population

71% Intent to Respond

- Above-average intent to respond.
- Highest percentage White.
- Highest percentage male.

Confidentiality Minded

15% of U.S. Population

63% Intent to Respond

- Most concerned answers will be used against them.
- Highest percentage foreign-born.
- Slightly below average intent to respond.

Head Nodders

9% of U.S. Population

60% Intent to Respond

- Most likely to respond "Yes" to all knowledge questions.
- Above-average percentage of foreign-born people.
- Highest percentage ages 18 to 34.

Wary Skeptics

14% of U.S. Population

59% Intent to Respond

- Lowest trust in government.
- Highest apathy about the census.
- High percentage of Black/African-Americans.

Disconnected Doubters

10% of U.S. Population

51% Intent to Respond

- Lowest intent to respond.
- Lowest frequency of internet use.
- Highest percentage age 65 or older.

Notes: (1) Intent to respond is based on the 2020 CBAMS Survey question, "If the census were held today, how likely would you be to fill out the census form? (a) Extremely likely, (b) Very likely, (c) Somewhat likely, (d) Not too likely, (e) Not at all likely." The percentages reported above in "Intent to Respond" reflect answers of "Extremely likely" or "Very likely" to respond.
(2) Percentage of U.S. Population reflects the weighted percentage of 2020 CBAMS Survey respondents in each mindset group.
(3) Due to rounding, population percentages do not add to 100%.

Figure 1. Overview of Mindsets (McGeeney et al 2019)

In order for the campaign to reach its utmost potential, the research team wanted to assign a mindset to *every household* in the U.S. This would theoretically allow micro-targeting of advertising and messages at the household level. To achieve this goal, the contractor proposed development of a "typing tool". Once the six interpretable mindsets were identified above, the next step would tie every household in the U.S. to the mindset that best fit its perception of the census. This process is referred to as 'typing' households.

A commercially available dataset, hereafter called as the National Household file (NHF), was procured by Y&R to link households included in the CBAMS sample with households in the U.S. population. The team was interested in exploring the quality of the predictions of the 2020 CBAMS mindsets based on the demographic characteristics available in the NHF containing over 250 million adult records and ultimately to households. The 2020 CBAMS survey variables used in determining the mindsets, in particular, answers to questions regarding attitudes and behavior (e.g. familiarity and knowledge about the census), are not present on the NHF although the dataset does contain over 500 variables that reflect, directly or indirectly, demographics, socioeconomic status, attitudes and behavior. The plan was to link the 2020 CBAMS survey records to the NHF and then use multinomial logistic regression with independent variables from the NHF to predict the probabilities of the mindsets. Next, the plan called for using the model-based estimates of the probabilities of the six mindsets for each adult household members to assign the most likely mindset for the household at each address.

The advantage of assigning a mindset to each address is that the campaign would be able to tailor the digital ads sent to the address to persuade the residents to respond to the Census. This approach had the potential of a digital ad prompting residents to respond to the Census via the web immediately after viewing an ad tailored to the viewer's mindset. In addition, the research team could aggregate information about mindsets to the level of the census tract and higher, which would aid in designing the delivery of messages through channels other than the Internet.

However, due to unanticipated delays in the fielding of the CBAMS, development of the mindsets, and computing resources needed to analyze the NHF, the research team lacked the time to develop the typing tool for households. Instead, an alternative method was developed at a more aggregated level, the tract. The approach relied on combining the mindset segmentation with an audience segmentation that partitioned the population by geography and demographics. To create tract-level estimates of the distribution of mindsets, the research team used the geographic and demographic information collected for all the CBAMS survey respondents to assign them to audience segments (Kulzick, R. et al. 2019).

Next, the team weighted the portion of the CBAMS sample that fell within each potential segment to mirror the demographic distribution of that segment. ACS five-year tract-level demographic estimates aggregated using population weights served as marginal totals for each potential segment. The team then used iterative proportional fitting to assign a weight to each CBAMS respondent such that the aggregated weights within each potential segment matched the corresponding total for that potential segment. Once the weights accurately reflected the potential segment's demographic distribution, the weighted average of the mindsets of respondents living in the segment produced the segment-level estimate of the distribution of mindsets for that potential segment.

There still is interest in the feasibility of implementing the originally proposed methodology for the typing tool for planning of future communications campaigns. This paper documents efforts to implement the originally proposed typing tool methodology. The primary research questions are:

(1) Which variables from the NHF were significant in predicting mindset assignment?
(2) How well can the NHF variables predict household-level assignment of psychographic "mindsets" for purposes of the 2020 Census Communication Campaign?
(3) How would using predicted probabilities of mindsets from a multinomial logistic regression model to assign a mindset to a household work in practice?

3. Linking to Build a Data File

3.1 National Household File (NHF)

The NHF that the U.S. Census Bureau acquired had 688 variables, 504 of which were individual level variables and 184 of which were household level variables. After internal U.S. Census Bureau processing that includes implementing standard criteria for variables to retain based on data quality measures such as the number of records with the variable missing, 573 variables were available for our analyses. Some examples of the available variables include individual socioeconomic characteristics (e.g. income, home ownership, education, race and ethnicity, language, children at home etc.), contextual information (i.e. neighborhood characteristics, crime statistics, home value, etc.) and individual attitudinal and behavior characteristics (online presence, employment, voting behavior, buying habits, etc.) and geographic information (e.g. county, city, state, zip code, etc.). The source of this NHF data is a combination of public-use data from the IRS, Social Security Administration, Bureau of Labor Statistics, and U.S. Department of Agriculture. In addition, the NHF has publicly available voting data from state election administrators, state licensing agencies, and data pulled from commercial consumer databases.

3.2 Strategy for Building Model to Predict Mindsets

The strategy for building a model involved using predictor variables obtained by matching the NHF records matched to CBAMS survey respondents. The research team would build a "typing tool" using a multinomial logistic regression model, or other classification technique, to predict the probability of each of the mindsets for each adult in the household at an address. Then the mindset assigned to the household would be the one with the highest estimated probability among all the probabilities assigned to the adults in the household at the address. Section 7 discusses this approach to assigning a mindset to a household.

The ultimate goal for the mindset typing tool is generate probability distributions across the mindsets that might best fit each household. The initial stage of the modeling focuses on socioeconomic characteristics as they are the closest one variables to the ones used in the 2020 CBAMS survey file. Some households' characteristics may suggest that they fit less neatly within a single mindset. These households will have their predicted probability distributed more evenly across multiple mindsets. Alternatively, households whose characteristics suggest that they will consistently fit within a single mindset will have a high predicted probability of fitting that mindset and small probabilities, if any, of fitting the other mindsets. For example, if the characteristics of a household based on income, race, age and education are closer to the characteristics of the Eager Engager mindset, we are hypothesizing that this same household will have a lower likelihood to belong to the rest of the mindsets such as the Head Nodders and the Weary Skeptics

Our strategy for building the multinomial logistic regression model has the four basic steps: These steps are similar to the original proposal:

1. Link the 2020 CBAMS survey person records to NHF person records using address and demographic variables
2. Build a multinomial logistic regression model to predict probabilities of mindsets using NHF variables
3. Score each NHF record with probabilities of mindsets
4. Assign the mindset with the highest probability to the address (household)

3.3 Linking Strategy

The procedure to link records in the 2020 CBAMS survey records with the NHF records required multiple steps. The two types of records did not contain exactly the same identifying information. The records for the NHF contained both address and name. However, the 2020 CBAMS survey records contained the sampled address but not the respondent's name. To prepare for linking, the addresses in both sources were linked to the Census Bureau's Master Address File for assignment of Master Address File Identification number (MAFID). Since the NHF records contained names and other demographic information, the NHF records were assigned Protected Identification Keys (PIKs). These PIKs are essentially encrypted Social Security Numbers or Individual Tax Identification Numbers, which are included when we use the term Social Security Numbers. When a data file with records for persons does not come with Social Security Numbers as in the case of the NHF, the Census Bureau uses its system to look up Social Security Numbers in Social

Security Administration files and encrypt them by assigning PIKs (Wagner and Layne 2014, Mulry and Keller 2018). For each project using a third-party file, the Census Bureau develops a unique identification number (UID) that corresponds to each PIK in the project files as an additional measure to avoid disclosure of personal information. Sometimes the Census Bureau's system fails to assign a PIK to a record. For example, 90.3% of the 2010 Census enumerations received a PIK from the Census Bureau's system, but only 97% of the enumerations had enough information for an attempt to assign a PIK (Wagner and Layne 2014). Evaluation studies have shown that missing date of birth in a record is highly correlated with the system not assigning a PIK. In addition, an incomplete or fake name in a record is highly correlated with a PIK not being assigned (Wagner and Layne, 2014; Mulrow *et al.* 2011).

The CBAMS-NHF linking procedure, summarized in Table 1, first linked the records from the two systems on MAFIDs and then identified all the UIDs associated with the MAFID. Next, the procedure collected the NHF records for the UIDs associated with each MAFID that linked to a CBAMS survey record. Within each MAFID, the procedure identified the NHF record, or records, that agreed on gender and had birth years that agreed or were within two years. If more than one NHF record met the criteria, the priority for breaking the tie was (1) birth years agree, (2) birth years differ by 1 year, and (3) birth years differ by 2 years.

The linking procedure found that approximately 15,000 of the 17,500 MAFIDs with responses in the 2020 CBAMS were present in the NHF. The next step that attempted to link records for individuals was able to find approximately 11,000 NHF records that "looked like" CBAMS respondents in that the records agreed on MAFID, gender and birth year with a tolerance of a difference of 2 years. Most of the linked records agreed on both gender and birth year. For an additional 200 CBAMS respondents, multiple individual NHF records tied for the preferred link. We did not include these 200 records in the analyses presented here since we needed more familiarity with the data to develop good criteria to use to break the ties. Table 1 summarizes the results.

Table 1. Summary of linking the 2020 CBAMS survey records to the NHF records

Linkage variables	*NHF and CBAMS linking process findings*	*Status*
MAFID	15,000 of the 17,500 MAFIDs with responses in the 2020 CBAMS were present in the NHF	Included in the analysis
MAFID, gender and birth year with a 2-year tolerance	11,000 of the 15,000 MAFIDs that linked had 1 NHF record that met the criteria	Included in the analysis
Multiple individual NHF records tied for the preferred link	200 of the 15,000 MAFIDs that linked had more than 1 NHF record that met the criteria	Not included in the analysis

Note: Numbers rounded for disclosure avoidance.

4. Development of Typing Tool – Feasibility Study

The first step in building a multinomial logistic regression model to predict the probabilities of the mindsets for the NHF records is variable reduction and variable selection, which may be part of the model fitting software. We used the Variable Selection Node in SAS Enterprise Miner 14.2 for variable reduction and the Regression Node for model fitting.

4.1 Variable Selection

For our initial attempt, we decided to make management of NHF variables easier by selecting 67 of 573 variables that our experience with the formation of the mindsets led us to believe had potential to be helpful in forming models to predict the mindsets. This consensus meeting helped us identify the variables that existed both in the 2020 CBAMS public opinion survey data and in the NHF file. These variables were mainly demographic characteristics: age and gender. As a next step, we also identified the variables related to civic engagement (e.g. online usage, voting behavior) and socioeconomic status (e.g. education, income, etc.) as well as other control variables such as rural areas, state, stratum and Census low response score (a publicly available metric indicating propensity to self-respond in the 2010 Census; see Erdman and Bates, 2017).

We began by using the R-square option in the Variable Selection Node for variable reduction. The Variable Selection Node also automatically groups levels of a categorical variable in an optimal manner for the Target (Dependent) variable. The minimum R-square to retain a variable is 0.005.

Next, we used the Regression Node to fit the logistic regression models for each pair of mindsets, which resulted in 15 models. We used the stepwise procedure to further reduce variables when fitting the pairwise logistic regression models for mindsets. The variables included in the final version of each pairwise logistic regression model had a p-value for its chi-square statistics that was less than 0.05, and the chi-square statistic for the model itself was less than 0.05.

We chose this approach because it is well suited to our goal of producing probabilities of each NHF record have each of the 6 mindsets as opposed to scoring each record with a mindset. Other methods of variable reduction such as Decision Trees, Neural Networks, and Random Forest are available in SAS Enterprise Miner 14.2. Our initial examination of these methods did not produce large differences in the variable reduction. However, we intend to revisit these methods as we refine our initial results presented in this paper.

4.2 Misclassification rates for pairwise logistic regression models

Although we started with 67 variables that our experience showed were related to the mindsets, we did not know which variables contained the type of information required to differentiate mindsets in a multinomial logistic regression model. All of the 67 variables were main effects; none were interactions. To answer this question, we fit 15 logistic regression models where the levels of the dependent variable were a pairs of mindsets. Our thought was that variables that produced a reasonably good fits for the pairwise models would have potential for contributing to the development of a model to predict the mindset for individual records. The dependent variable of the pairwise logistic regression models is one of the pairs of mindsets compared to each other. We used 67 independent variables and let a stepwise procedure select the model that produced the best estimates based on the significance of the estimated coefficients of the selected variables.

The classification rates for the 15 models shown in Table 2 have misclassification rates ranging from 0.148 to 0.370. The higher this misclassification rate is, the less the certainty that we can distinguish one mindset from the other mindset in the pairwise model. The models appear to break into three groups with respect to misclassification rates: 5 models with rates between 0.148 and 0.181 (low misclassification), 6 models with rates between 0.231 and 0.277 (low medium misclassification), and the remaining 4 models with rates between 0.327 and 0.370 (high medium misclassification).

The four models in the high medium misclassification rates group apparently have the pairs of mindsets that are the most difficult to distinguish (based on the sorting of the misclassification rate number): Confidentiality Minded vs. Wary Skeptics, Eager Engagers vs. Fence Sitters, Confidentiality Minded vs. Head Nodders, and Head Nodders vs. Wary Skeptics. This means, for example, that when we compare the Confidentiality Minded mindset assignments at the individual level, it is harder to differentiate this particular mindset from the Wary Skeptics and the Head Nodders than we saw with the pairs in the low and medium groups. The household level available data does not allow us to distinguish as clearly between these mindsets. In the other extreme, according to the misclassification rate, the pairwise logistic regression models indicate that the mindset of Eager Engagers is distinguishable from the Disconnected Doubters. This is good news since according to Figure 1 the Disconnected Doubters is the mindset least likely to participate in the Census and is clearly the opposite of the Eager Engagers mindset that has the largest intention to participate in the Census. The Fence Sitters mindset is the second group with the highest intention to participate in the Census (see Figure 1) and not surprisingly, this mindset is distinguishable from the Disconnected Doubters, which has the lowest intention to participate compared to the rest of the mindsets.

Table 2. Misclassification rates for 15 logistic regression models where the two levels of the dependent variable are mindsets and the variables are selected using a stepwise procedure, sorted from smallest to largest misclassification rate

Mindset pair	*Misclassification rate*
Eager Engagers vs Head Nodders	0.148
Eager Engagers vs Discontented Doubters	0.148
Fence Sitters vs Discontented Doubters	0.160
Fence Sitters vs Head Nodders	0.180
Head Nodders vs Discontented Doubters	0.181
Confidentiality Minded vs Fence Sitters	0.231
Fence Sitters vs Wary Skeptics	0.233
Confidentiality Minded vs Discontented Doubters	0.239
Wary Skeptics vs Discontented Doubters	0.243
Eager Engagers vs Confidentiality Minded	0.246
Eager Engagers vs Wary Skeptics	0.277
Head Nodders vs Wary Skeptics	0.327
Confidentiality Minded vs Head Nodders	0.336
Eager Engagers vs Fence Sitters	0.348
Confidentiality Minded vs Wary Skeptics	0.370

5. Important variables for distinguishing each mindset

Since each mindset was one of the levels in 5 logistic regression models, the number of times that the stepwise procedure selected the variable indicates how important the variable is in distinguishing the mindset. Table 3 shows the number of times a variable was selected when a mindset was one of the levels in the logistic regression models.
The variable Head of Household Salary appeared in every one of the pairwise models which leads us to conclude it is likely the most important variables overall. The variable State appears to be the second most important variable since it was included in 13 models. Stratum was the third variable appearing as in the Eager Engagers, Fence Sitters and Head Nodders models. Focusing on the columns in Table 3, we also notice that the Eager Engagers and Disconnected Doubters models had the largest concentration of significant variables (6 out of 11). Interestingly, there are some differences when comparing the Eager Engagers and Disconnected Doubter models. For instance, the voting variables are selected only for the Eager Engagers mindset but not for the Disconnected Doubters. The variable of Percent Rural Population in Tract in the 2010 census, Household Size, Birth Year and Facebook Users were selected only for the Disconnected Doubters and not for the Eager Engagers. Across the models, the College Graduate variable was selected only for models where one of the mindsets was the Eager Engager or the Wary Skeptic.

Each of the 5 pairwise logistic regression models where the Confidentiality Minded mindset was one of the levels included the variables Head of Household Salary and State. No other variables appeared in 3 or more of these models.

Table 3. The number of times a variable was selected when a mindset was one of the levels for the dependent variable in the pairwise logistic regression models with a lower bound of 3 models, sorted by number of times

Variable	*Eager Engagers*	*Conf Minded*	*Fence Sitters*	*Head Nodders*	*Wary Skeptics*	*Disconnected Doubters*
Head of household salary	5	5	5	5	5	5
State	5	5	4	4	5	3
Stratum	4		3	3		
Percent rural population in tract in 2010 Census				3		3
Household size			4			3
College graduate	3				3	
Birth year						5
Voted in 2014 election	5					
Voted in 2006 election	3					
Low Response Score of tract			3			
Household on Facebook						3

The Eager Engager mindset had 6 variables that appeared influential in distinguishing it from the other mindsets, which included Head of Household Salary and State. In addition, a binary variable reflecting level of education, named College Graduate, appeared in 3 of the logistic regression models where Eager Engager was one of the levels. Two binary variables that reflect civic engagement in the form of voting in non-Presidential election years appeared influential. These variables are: (1) Voted in 2014 Election, which was selected for 5 models, and (2) Voted in 2006 Election, which was selected for 3 models. Four of the pairwise logistic models included the Stratum variable which reflects the race/Hispanic ethnicity of the tract where the address is located and the type of contact strategy used by the American Community Survey in the tract.

The Discontented Doubter mindset also had 6 variables that appeared influential in distinguishing it in the pairwise logistic regression. These included the variables Head of Household Salary in 5 models and State in 3 models. Also appearing in 3 models were the variables Household Size, Household on Facebook, and Percent Rural Population in Tract in 2010 Census. The variable Birth Year appeared in all 5 logistic regression models that included Disconnect Doubter as one of the levels in the dependent variable. Interestingly, Disconnected Doubter is the only mindset where age in the form of Birth Year appeared important in distinguishing it from the other mindsets.

The variables that appear important in distinguishing the Fence Sitter mindset include some that already have been mentioned and a new one. These included the Head of Household Salary in 5 models and State in 3 models. Also appearing in 3 models were the variables Household Size in 4 models and Stratum in 3 models. The new variable is the Low Response Score of the tract, which is based on a model that employs ACS data to reflect the level of difficulty in obtaining responses to surveys in the tract.

The models for the remaining 2 mindsets, Head Nodder and Wary Skeptic, indicate that the variables important in distinguishing them overlap with those identified for other mindsets. For the Head Nodder, Head of Household Salary appears in 5 models, State in 4 models, and Stratum in 3 models. Three models included the Low Response Score for the tract. For the Wary Skeptic, Head of Household Salary appears in 5 models, State in 5 models, and the binary variable College Graduate in 3 models.

6. Model to predict Mindsets

Next, we focused on fitting a multinomial logistic regression model for the categorical Mindset variable with six levels, one for each mindset. We started with the 35 variables that were selected for at least one of the pairwise models discussed in Section 5 and then used a stepwise selection to fit the mindset model. Table 4 shows the selected variables.

Each of the 19 variables selected by the stepwise procedure for the multinomial logistic regression model to predict the six mindsets had a p-value for its chi-square statistics that was less than 0.05, and the chi-square statistic for the model itself was less than 0.05. However, the misclassification rate is 0.58 caused by having almost twice as many Fence Sitters, the largest group, as there should be.

The Variable Selection Node grouped some of the variables for each of the pairwise models. The Variable Selection Node did not group exactly the same variables for each model, and the groupings were not necessarily the same for all the models where the stepwise procedure selected a particular. The candidates for the Mindset model included 5 variables that used the groupings formed to fit the model for the Eager Engagers vs. Fence Sitters because these are the two largest mindsets. The groupings for two other candidate variables were those formed for the Fence Sitters vs. Wary Skeptics model since these mindsets rank first and third in size among the mindsets.

Table 4. Variables selected for multinomial logistic regression model to predict Mindsets

Birth year
College graduate
Gender
Head of household salary (grouped)
Someone in household voted in 2016
Household size (grouped)
Household made health institution purchase
Household has online purchaser
Household has a retired person
Homeowner
Length of residence
Low Response Score of tract
Household on Facebook
Percent rural population in tract in 2010 Census
State (grouped)
Stratum (grouped)
Voted in 2006
Voted in 2010
Voted in 2014

7. Discussion of assigning Mindsets to households

In the previous sections, we have shown that the NHF includes variables that contain the information needed to differentiate the between the 6 mindsets uncovered by using the responses to attitudinal questions in the 2020 CBAMS sample survey. At a minimum, these variables appear effective in differentiating between the mindsets on a pairwise basis since we observed misclassification rates for the pairwise models ranging from 0.148 to 0.370, In addition, we demonstrated the feasibility of fitting a six-category multinomial model to predict the mindsets. We plan to continue refining our variables and the models to predict the probabilities of the mindsets for the individual records in the NHF.

Next, we turn our attention to the assignment of the mindset with the highest probability to the household. When assigning a mindset to a household, one would like for the probability that at least one person in the household has that mindset to be 0.5 or greater. However, it is not clear that one can build a multinomial logistic regression model that produces an estimated probability of 0.5 or greater for at least one member for every household, particularly the small households with only one or two adults. The correct classification rates for the individual mindsets may be helpful when all estimated mindset probabilities of all the household members are less than 0.5.

Table 5 illustrates the complications that can be encountered when assigning predicted probabilities. It is hard to find variables that will produce a record with a probability of 0.5 or greater for one person in every household. Table 5 shows the probabilities of mindsets for 3 persons predicted from a multinomial logistic regression model fit using only 2020 CBAMS survey data for a subpopulation and then used to assign mindsets at the person level and household level based on the predicted probabilities of mindsets. The probabilities of the all mindsets predicted for Person 1 and Person 2 are equal to 0.28 or lower and therefore, definitely not greater than 0.5. However, the probability of the Confidentiality Minded assigned to Person 3 equals 0.58, slightly greater than 0.5. When we consider a household that contains all 3 persons, the proposed rule assigns the Confidentiality Mindset to the household since it is the largest for any mindset for all 3 persons. Calculating the probability of at least one person in the households has a probability greater than 0.50 of having the Confidentiality Minded mindset equals the sum of three probabilities, the probability all 3 persons are Confidentiality minded, the probability exactly 2 persons have the Confidentiality mindset, and the probability exactly 1 person has the Confidentiality Mindset. For the household with Persons 1, 2, and 3, the probability that at least one person has the Confidentiality Mindset is 0.76. In this case, the strategy of assigning the household the mindset with the highest probability over all the household members works well.

Table 5. Probabilities of mindsets for 3 persons predicted from a multinomial logistic regression model fit with 2020 CBAMS sample survey data for a subpopulation. Shading identifies the mindset with highest probability for each person.

	Eager Engager	***Conf Minded***	***Fence Sitter***	***Head Nodder***	***Wary Skeptic***	***Disconn Doubter***
Person 1	0.09	**0.28**	0.09	0.13	0.21	0.20
Person 2	0.19	0.23	**0.24**	0.11	0.20	0.03
Person 3	0.02	**0.58**	0.10	0.12	0.15	0.03
Household composition						
<u>Observed</u>		***<u>Highest</u>***	***<u>Probabilities</u>***			
Persons 1,2,3		Conf Minded	Prob(Conf Minded for Person 3) = 0.58 Prob(at least 1 person is Conf Minded) = 0.76			
Persons 1,2		Conf Minded	Prob(Conf Minded for Person 1) = 0.28 Prob(at least 1 person is Conf Minded) = 0.44			
Person 1		Conf Minded	Prob(Conf Minded for Person 1) = 0.28 Prob(at least 1 person is Conf Minded) = 0.28			

On the other hand, if the household has only Person 1 and Person 2, none of the mindsets has a predicted probability greater than 0.28. Person 1 and Person 2 each have 3 mindsets with probabilities ranging from 0.20 to 0.28. The mindset with the highest probability of 0.28 is Confidentiality Minded for Person 1. The probability that at least one of the residents has the Confidentiality Minded mindset is 0.44. The reliability of assigning the Confidentiality Minded mindset to the household when the probability that at least 1 person has that mindset is lower than 0.50 seems low. However, if the researcher knows the estimated probability comes from a multinomial logistic regression model that produced a correct classification rate greater than 0.5 for the Confidentiality Minded, then assigning the Confidentiality Mindset when it had the highest estimated probability would be less of a concern. The same rationale applies to 1-person households such as the one shown in Table 5 where Person 1 is the only household member.

However, situations such as the household with Person 1 and Person 2 and the household with only Person 1 but the multinomial logistic regression model producing the estimated probabilities did not have a correct classification rate greater than 0.5 may call for additional steps in the assignment of a mindset to the household. Some examples of such steps include a sequential set of models, several models fit for specific race/Hispanic ethnicity groups or geographic areas rather than the entire U.S., and a set of models based on the knowledge of the similarity between communication materials prepared for mindsets. An example of a sequential set of models is to have a model for the entire U.S. but to use additional models such as the pairwise models discussed in the previous section to break ties when needed. Another approach would be to a fit 3-level multinomial logistic regression models with combinations of the mindsets with the highest probabilities for use in assigning a mindset to the household.

8. Conclusion/Lessons Learned

Our research illustrates the feasibility of developing a multinomial logistic regression model for assigning mindsets to records for individuals from a third-party dataset (the NHF) with over 250 million adult records and over 500 variables. The goal of our research is to demonstrate the feasibility of expanding results from a national sample survey to the entire U.S. population by leveraging a “typing tool” and Big Data. Our challenges included data matching difficulties, variable selection and reduction choices, and unexpected constraints in modeling a six-category dependent (target) variable. Our next steps include exploring whether we can refine and improve the model. The pairwise logistic models illustrate that there are variables important for distinguishing the mindsets from each other. Therefore, our goal is find a better model that reduces the misclassification rate or, at a minimum, has most misclassifications be between mindsets that are ‘close’ in attitudes.

We also would like to conduct a practical evaluation of the model that would involve comparing the mindsets assigned to households at addresses in the NHF to the mindsets that correspond to mindsets used to messages deliver messages to the households during the 2020 Census. Because the 2020 campaign research team ultimately assigned mindsets at the tract-level, and not household level, this will necessitate us aggregating household-level mindset to produce tract-level mindset distributions

References

AAPOR (2016), “Standard Definitions Report, Ninth Edition,” AAPOR. Available at https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf (Accessed September 2019).

Bates, N. (2017), “The Morris Hansen Lecture. Hard-to-Survey Populations and the U.S. Census: Making Use of Social Marketing Campaigns,*” Journal of Official Statistics*, Vol. 33, No. 4, 2017, pp. 873–885.

Erdman, C. and Bates, N. (2017), “The Low Response Score (LRS): A Metric to Locate, Predict, and Manage Hard-to-Survey Populations”. *Public Opinion Quarterly*, 81, 1, pp. 144–156. Available at https://doi.org/10.1093/poq/nfw040.

Evans, W., A. Douglas, R. Datta, and T. Yan (2014), “Use of paid media to encourage 2010 Census participation among the hard to count,” in *Hard-to-Survey Populations*, eds. R. Tourangeau, B. Edwards, T. P. Johnson, K. Wolter, and N. Bates, pp. 519-540, Cambridge, UK: Cambridge University Press.

Kulzick,R., Kail, L., Mullenax, S., Kriz, B., Shang, H., Walejko, G., Vines, M., Bates, N. Schied, S. and García Trejo, Y. (2019). *2020 Census Predictive Models and Audience Segmentation Report*. 2020 Census Research Memorandum Series. Washington, DC: U.S. Census Bureau. Available at: https://www.census.gov/programs-surveys/decennial-census/2020-census/research-testing/communications-research/2020-tract-segments.html (Accessed September 2019).

McGeeney, K., Kriz, B., Mullenax, S., Kail, L., Walejko, G., Vines, M., Bates, N. and García Trejo, Y. (2019), *2020 Census Barriers, Attitudes, and Motivators Study Survey Report*. 2020 Census Research Memorandum Series. Available at https://www2.census.gov/programs-surveys/decennial/2020/program-management/final-analysis-reports/2020-report-cbams-study-survey.pdf (Accessed September 2019).

Mulrow, E., A. Mushta, S. Pramanik, and A. Fontes (2011), Assessment of the U.S. Census Bureau’s Person Identification Validation System. Report for the U.S. Census Bureau. Chicago, IL: NORC. Available at: http://www.norc.org/PDFs/May%202011%20Personal%20Validation%20and%20Entity%20Resolution%20Conference/PVS%20Assessment%20Report%20FINAL%20JULY%202011.pdf (accessed September 2019).

Mulry, M. H. and A. Keller (2017), "Comparison of 2010 Census Nonresponse Followup Proxy Responses with Administrative Records Using Census Coverage Measurement Results." *Journal of Official Statistics.* 33(2). pp. 455–475.
DOI: https://doi.org/10.1515/jos-2017-0022

Wagner, D. and M. Layne (2014), “The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications.” CARRA Working Paper Series. Working Paper #2014-01. Washington, DC: U.S. Census Bureau. Available at: https://www.census.gov/library/working-papers/2014/adrm/carra-wp-2014-01.html (accessed September 2019).

Wedel, M. and W. A. Kamakura (2012). *Market Segmentation: Conceptual and Methodological Foundations*, New York, NY: Springer Science and Business Media.

Williams, J. D., N. Bates, M. A. Lotti, and M. J. Wroblewski (2014), “Marketing the 2010 Census: Meeting the Challenges of Persuasion in the Largest-Ever Social Marketing Campaign,” in *The Handbook of Persuasion and Marketing*, ed. D. W. Stewart, pp. 117–154, Santa Barbara, CA: Praeger.

Xiong, H. and Z. Li (2013), “Clustering Validation Measures,” in *Data Clustering Algorithms and Applications*, eds: C.C. Aggarwal and C. K. Reddy, pp.571–605, Boca Raton, FL: Chapman and Hall/CRC (Taylor and Francis Group).