Analyzing Tradeoff Between Administrative Records Enumeration and Count Imputation

Andrew Keller[1]
U.S. Census Bureau, 4600 Silver Hill Rd., Washington, DC 20233

Abstract

After completing field operations for a decennial census, there are some addresses with an unresolved status (occupied, vacant, and delete) and household count. Since 1960, after decennial census field operations, a count imputation procedure has filled in missing household status and size for the small proportion of addresses where this information has been unknown. For the 2020 Census, the U.S. Census Bureau is researching the use of administrative records to provide a status and count for some nonresponding addresses instead of imputing a status and count. This paper discusses tradeoffs between enumerating the unresolved addresses with administrative records versus imputing the missing status and count. Specifically, we analyze the extent to which administrative records enumeration is preferable before count imputation is applied. The goal is to establish a quality threshold for administrative records to determine when count imputation is a better alternative to administrative records enumeration.

Key Words: Census, Count Imputation, Administrative Records

1. Introduction

The Census Bureau has researched fundamental changes to the design and implementation of the 2020 Census. One major research area noted in the 2020 Operational Plan (U.S. Census Bureau 2018a) is to incorporate administrative records (AR) into the census design. The U.S. Census Bureau has proposed using administrative records for updating the address frame, advertising, validating respondent addresses for Internet responses to prevent fraud, and reducing contacts in the Nonresponse Followup (NRFU) operation. This research looks at using administrative records in lieu of imputation procedures.

Section 2 provides background concerning the count imputation procedure. Section 3 discusses how administrative records are used for enumeration. Section 4 introduces the parameters for performing the tradeoff between count imputation and using administrative records for enumeration. Section 5 and 6 show results and discuss takeaways from the research.

[1] Any views expressed are those of the authors and not those of the U.S. Census Bureau. The Census Bureau's Disclosure Review Board and Disclosure Avoidance Officers have reviewed this data product for unauthorized disclosure of confidential information and have approved the disclosure avoidance practices applied to this release. (CBDRB-FY19-ACSO002-B0020)

2. Count Imputation

Census operations attempt to obtain complete counts of the population and housing units. To do this, each address must be assigned a final housing unit status of occupied, vacant, or non-existent (also known as a delete). If the status is occupied, then the housing unit record must also have a population count greater than zero. At the end of the census data collection operations, some addresses lack a housing unit status or population count. Depending on what is known about the address, records needing imputation are classified into three categories. Status imputation cases occur when there is not enough information to know whether the address is an occupied, vacant, or non-existent housing unit. Occupancy imputation cases occur when the address is an existing housing unit but it is unclear whether it is occupied or vacant. Household size imputation cases are known to be occupied housing units but are missing a population count.

Thus, count imputation has two functions:

1) Fill in missing housing unit status (occupied, vacant, or non-existent).
2) Fill in missing population counts for housing units that are known to be occupied but household size is unknown, or imputed to be occupied.

Due to the extensive visit protocol for the 2010 Nonresponse Followup (NRFU) operation, the 2010 Census had a very small rate of count imputation (0.38%).

The count imputation method used in this research is the nearest-neighbor hot deck. We create imputation cells by combining seven variables that are correlated with a) the propensity to be unresolved, and b) the housing unit status (occupied, vacant, non-existent) in the 2010 Census data. The seven variables used are:

- Nearest-Neighbor Household Type - the household type of the record’s nearest resolved neighbor (e.g. occupied, vacant, non-existent).
- Master Address File (MAF) Unit Status - valid living quarters or not (e.g., demolished, delete, duplicate).
- MAF X-Type Flag - classifies the record as likely delete, likely vacant, or other.
- Spring Delivery Sequence File (DSF) Flag - classifies the record as residential or other (e.g., commercial or not on the DSF).
- NRFU Proxy Type - classifies units visited in NRFU as having an unknown proxy respondent or other (e.g., no proxy, household member).
- Undeliverable as Addressed (UAA) Reason Code - classifies the record into one of three categories: No such number, all other UAA codes, or no UAA code (includes addresses with no mailing).
- Count of Administrative Records - population count in address according to administrative records (0 – 9+).

3. Administrative Records Modeling and Enumeration

Previous research has developed methods to combine and use several administrative sources to identify occupied, vacant, and non-existent units prior to or after minimal NRFU fieldwork, thus reducing the number of enumerator visits (Keller et al., 2018). This allows resources to focus on units where administrative data are unreliable or unavailable. In this paper, we repurpose the approach of using administrative records to classify units. The context occurs in the data-processing phase. That is, we determine when to use administrative records as opposed to imputation.

3.1 Administrative Records Enumeration Distance Thresholds

To identify vacant units with AR, we developed a multinomial logit model which predicted the probability that an AR address would have been enumerated as vacant during the 2010 Census. Keller et al. (2018) provide more discussion of the vacancy model. Independent variables in the model included variables indicating whether the census mailings could be delivered to the address and whether the AR sources indicated anyone lived at the address. The dependent variable had three probabilities associated with each address in the NRFU universe:

- occupied
- vacant, or
- delete (i.e., not a HU).

We defined a Euclidian vacant distance function for AR Vacant identification as:

$$d_{AR_{Vac}} = \sqrt{(1 - p_{vacant})^2 + \left(0 - p_{occupied}\right)^2}$$

The formula shows that cases with the smallest distance were those with the highest vacant probability and lowest occupied probability. Starting with the smallest vacant distance, AR Vacant cases were identified by allowing for increased vacant distance values up to a threshold. This threshold was based on analysis of 2010 Census NRFU data.

We defined a Euclidian delete distance function for AR Delete identification as:

$$d_{AR_{Del}} = \sqrt{(1 - p_{delete})^2 + \left(0 - p_{occupied}\right)^2}$$

The formula shows that cases with the smallest distance were those with the highest delete probability and lowest occupied probability. Starting with the smallest delete distance, AR Delete cases were identified by allowing for increased delete distance values up to a threshold. This threshold was based on analysis of 2010 Census NRFU data.

Two models were developed to identify AR Occupied units: a person-place model and a household (HH) composition model. Independent variables in the occupied models included variables indicating which AR sources placed people at the address and whether these people were found at a different address in the AR sources. The person-place model predicted the probability that an AR person would be enumerated at the sample address if fieldwork was conducted. The dependent variable was whether the AR person was at the address in the 2010 Census. The HH composition model predicted the probability that the sample address would have the same HH composition determined by NRFU fieldwork as its pre-identified AR HH composition. HH composition is defined by the number of adults in the unit and the absence or presence of children. The dependent variable was the 2010 Census HH composition. Keller et al. (2018) provide more discussion of the person-place and household composition models.

Similar to AR Vacant and AR Delete, we defined a Euclidian occupied distance function for AR Occupied identification as:

$$d_{AR\_Occ} = \sqrt{\left(1 - p_{person-place}\right)^2 + \left(1 - p_{HH\ composition}\right)^2}$$

The formula shows that cases with the smallest occupied distance were those where the person-place probability was closest to one and the household composition probability was closest to one (i.e. the (1,1) point). Starting with the smallest occupied distance, AR Occupied cases were identified by allowing for increased occupied distance values up to an occupied threshold.

4. Administrative Records Simulation

At the beginning of the NRFU operation, vacant, delete, and occupied distance thresholds are selected so a small amount of cases are identified as AR Vacant, AR Delete, and AR Occupied. For this simulation, the distance thresholds are 0.330 for AR Vacant and AR Delete and 0.685 for AR Occupied.

The research goal of this paper is to understand when to use count imputation as opposed to more administrative records enumeration. To begin, using the 2010 Census data, we simulate unresolved records. In essence, for a selected set of addresses we erase the observed response and treat it as missing. We call this scheme a truth deck. This allows us to compare simulated results against the 2010 reported responses.

4.1 Truth Deck Formulation

The truth deck methodology uses AR data, operational, and block-group level variables to predict the propensity for housing unit status, occupancy, and household size to be missing or unresolved. It is documented in Williams (2005). The first step is to create three models, one to predict status imputation, another to predict occupancy imputation, and the last to predict household size imputation. The status imputation model is fit over all addresses. The occupancy imputation model is fit over all occupied or vacant housing units. The household size imputation model is fit over only occupied addresses.

The second step uses the missingness probabilities from each model. We use it to flag addresses as missing or non-missing. We use each of the three models to predict the type of missingness (status, occupancy, or household size) probability for all addresses in the 2010 Census universe.

We first compare a random uniform draw with the status imputation predicted probability. If the random draw is less than the status imputation predicted probability, the case is flagged for status imputation. For all remaining cases not flagged as status imputations, we attempt to flag them as occupancy imputations. Address records with a status of delete are ineligible to be flagged as occupancy imputations. If the random draw is less than the occupancy imputation predicted probability, the case is flagged for occupancy imputation. For all remaining cases not flagged as status or occupancy imputations, we attempt to flag them as household size imputations. Only address records with an occupied status are eligible to be flagged as household size imputations. Similar to other imputation types, if the random draw is less than the household size imputation predicted probability, the case is flagged for household size imputation. Note that the predicted probabilities that are compared to the random draw can be multiplied to achieve a higher or lower imputation rate than was seen in the 2010 Census.

For this research, we developed three truth decks. The first truth deck had 1,210,000 cases modeled as unresolved. After the missing cases were determined, those cases with administrative records within the initial distance threshold were treated as resolved. This resulted in 1,000,000 unresolved cases. The observed distribution among those missing cases was 82.6% occupied, 10.7% vacant, and 6.7% delete. The second truth deck had a similar observed distribution – the difference was that 1,329,000 cases are missing as opposed to 1,000,000. The third truth deck had a different observed distribution among the missing cases - 77.4% occupied, 14.0% vacant, and 8.6% delete. Differences in the truth decks were created by including different independent variables for the model fitting and multiplying the predicted probabilities.

5. Results

The research explores when to use AR data versus count imputation to resolve cases. We cannot use AR for all addresses as some do not have AR associated with them. In other words, while count imputation is necessary for some portion of this unresolved universe, we are trying to ascertain the point at which we need to stop using administrative records and instead use count imputation because the AR data yields less desirable results than count imputation. A major benefit of using a roster from AR as opposed to count imputation is that the AR has known characteristics for the household roster. These include age, sex, Hispanic origin, race, and relationship to householder. If count imputation were to be used, all characteristics for the household roster would need to be imputed.

5.1 Tradeoff Scenarios

To begin, we construct six scenarios. The first scenario only uses count imputation. The remaining five scenarios use progressively more generous distance thresholds which allow for more AR usage. Table 1 shows a column for the occupied distance threshold and vacant/delete distance threshold. Then for each truth deck, we show the percentage of AR used for the given distance threshold scenario. Note that, depending on the truth deck, a different percentage of AR can be used for the same distance thresholds.

Scenario 0 shows the simulation in which only count imputation is used. The occupied distance threshold of 0.685 and vacant/delete distance threshold of 0.330 represent the distance thresholds used at the beginning of the NRFU operation that identify the highest quality AR cases. All cases within these thresholds are resolved either through census operations or administrative records. Beyond the Scenario 0, we consider larger thresholds.

Scenario 1 has an occupied distance threshold of 0.753 and vacant/delete distance threshold of 0.398. With respect to Truth Deck 1, using these thresholds resolves approximately 5% of the unresolved workload. For the remaining 95% of unresolved cases, the count imputation method is performed. However, with respect to Truth Deck 3, using these thresholds resolves approximately 6% of the unresolved cases with AR.

Table 1: Distance Threshold Scenarios

Scenario Name	Occ Distance	Vac/Del Distance	% AR Used		
			Truth Deck 1	Truth Deck 2	Truth Deck 3
Scenario 0	0.685	0.330	0.0%	0.0%	0.0%
Scenario 1	0.753	0.398	5.0%	5.0%	6.0%
Scenario 2	0.828	0.473	10.0%	10.0%	12.0%
Scenario 3	0.911	0.556	14.8%	14.9%	17.9%
Scenario 4	1.016	0.661	20.5%	20.5%	24.7%
Scenario 5	1.333	0.978	24.5%	24.5%	28.8%

5.2 Distributional Analysis

We compare the simulated statuses against the true response in our truth deck. We are first interested in the address status - occupied, vacant, or delete. We start out by looking at the distributional accuracy of combining AR enumeration and count imputation seen in the various scenarios. Table 2 displays the distribution of occupied, vacant, and delete status for the truth and the various scenarios. The top row is the observed distribution, i.e., the distribution that was observed for these units in the 2010 Census.

Table 2: Status Distributions of Scenarios

Scenario	Truth Deck 1			Truth Deck 2			Truth Deck 3		
Name	% Occ	% Vac	% Del	% Occ	% Vac	% Del	% Occ	% Vac	% Del
Observed	82.6%	10.7%	6.7%	82.7%	10.7%	6.6%	77.4%	14.0%	8.6%
Scenario 0	83.1%	10.7%	6.2%	83.0%	10.7%	6.2%	78.1%	13.9%	8.0%
Scenario 1	83.0%	11.0%	6.1%	82.9%	11.0%	6.1%	77.8%	14.3%	7.9%
Scenario 2	82.7%	11.2%	6.0%	82.7%	11.3%	6.1%	77.5%	14.7%	7.8%
Scenario 3	82.4%	11.5%	6.0%	82.4%	11.6%	6.1%	77.1%	15.1%	7.8%
Scenario 4	82.1%	11.7%	6.2%	82.1%	11.7%	6.2%	76.6%	15.3%	8.1%
Scenario 5	83.2%	10.9%	6.0%	83.1%	10.9%	6.0%	78.0%	14.4%	7.7%

For the approximately one million unresolved cases in Truth Deck 1, the observed distribution in the 2010 Census was 82.6% occupied, 10.7% vacant, and 6.7% delete. For each truth deck, we evaluate the accuracy of the scenarios against this metric. For Truth Deck 1, Scenario 0 shows, if we were to disregard an AR cases and solely perform count imputation, the distribution would be 83.1% occupied, 10.7% vacant, and 6.2% delete. In this scenario, we simulated too many occupied units. As we go down the rows for Truth Deck 1, it appears that the observed occupied distribution is achieved somewhere between using Scenario 2 and Scenario 3. That is, the simulated distribution of occupied units matches that of the observed distribution of occupied units. That is also true for Truth Deck 2 and Truth Deck 3. For Scenario 4, too few cases are occupied and for Scenario 5, too many cases are occupied.

For each truth deck, Scenario 0 results in the nearly same percentage of vacant cases as the observed distribution. The exception is Truth Deck 3 – however it is only 0.1% lower than the observed distribution. If matching the vacant distribution was the primary concern, this would appear to be a promising scenario. However, Keller and Fox (2012) estimated an undercount of approximately 750,000 vacant housing units in the 2010 Census. As a result, more appropriate scenarios are those which have a vacant distribution which exceeds what was observed in the 2010 Census.

5.3 Individual-Level Analysis Using Receiver Operating Characteristic Distance

We analyze the proposed distance thresholds by determining the threshold that minimizes the Euclidean receiver operating characteristic (ROC) distance described in Metz (1978). The ROC curve plots the false positive rate (FPR) on the horizontal axis against the true positive rate (TPR) on the vertical axis. Each scenario is a modeling strategy with a mix of AR and count imputation use and has a (x_m, y_m) point associated with it. A perfect strategy would have a FPR=0 and TPR=1, i.e. $\left(x_p = 0, y_p = 1\right)$.

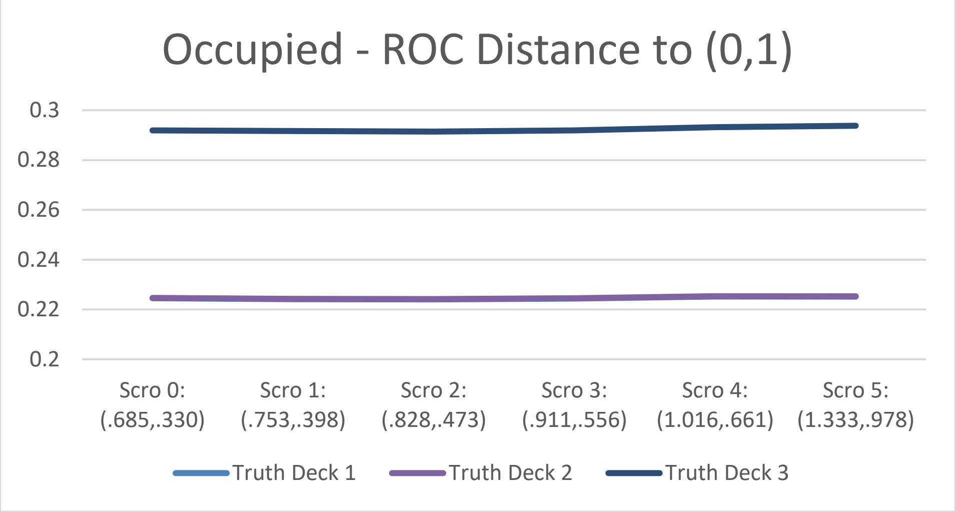
We would like to minimize the space between the point associated with the modeling strategy and the perfect strategy. The Euclidean ROC distance between the perfect strategy and the modeling strategy is defined as:

$$d_m = \sqrt{\left(x_p - x_m\right)^2 + \left(y_p - y_m\right)^2} = \sqrt{(0 - x_m)^2 + (1 - y_m)^2} = \sqrt{x_m^2 + (1 - y_m)^2}$$

For each status (occupied, vacant, delete) we calculate a ROC distance value for each scenario for each truth deck. Chart 1 plots the ROC distance values for the three truth decks for occupied classification. First, note that the line plots associated with Truth Deck 1 and Truth Deck 2 overlap each other while the line plot associated with Truth Deck 3 is different. With respect to Truth Deck 1 and Truth Deck 2, Table 2 shows that the observed distribution of occupied, vacant, and delete cases is the same even though the magnitude of unresolved cases is different. Interestingly, Chart 1 shows that the ROC distance measures overlap each other. This shows that magnitude of missingness may not make a difference with respect to deciding upon a distance threshold for when to stop using AR for enumeration.

Notice the relative flatness of each line plot. This means that the ROC distances are largely the same regardless of the scenario. Since the ROC distances are largely the same, it would justify using more AR given the benefit of not having to impute characteristics when completing AR enumerations as opposed to count imputations. Alternatively, it could also suggest using AR does not show much improvement over count imputation. Given those two arguments, it is important to restate that using AR instead of count imputation allows us to use AR to assign characteristics as opposed to performing characteristic imputation. That being said, all lines slightly go up after the (0.911, 0.556) distance threshold, meaning that the optimal distance threshold is probably there.

Chart 1: Occupied Distance Plots Associated with Perfect Classification Strategy



Charts 2 and 3 plot the ROC distance values for the three truth decks for vacant and delete classifications respectively. For both plots, the lines are less straight, indicating there are more distinct quality differences over the various scenarios. With respect to Chart 2, it appears that ROC distance is minimized at the (1.016, 0.661) distance threshold. However, we should recall that vacancy was undercovered in the 2010 Census, so this may not be optimal in terms of accuracy. With respect to Chart 3, the line plots appear to be flipped from the Chart 2 line plots. The ROC distance is maximized at the (1.016, 0.661) distance threshold while the minimum ROC distances appear to be for no AR use (0.685, 0.330) or maximum AR use (1.333, 0.978).

Chart 2: Vacant Distance Plots Associated with Perfect Classification Strategy

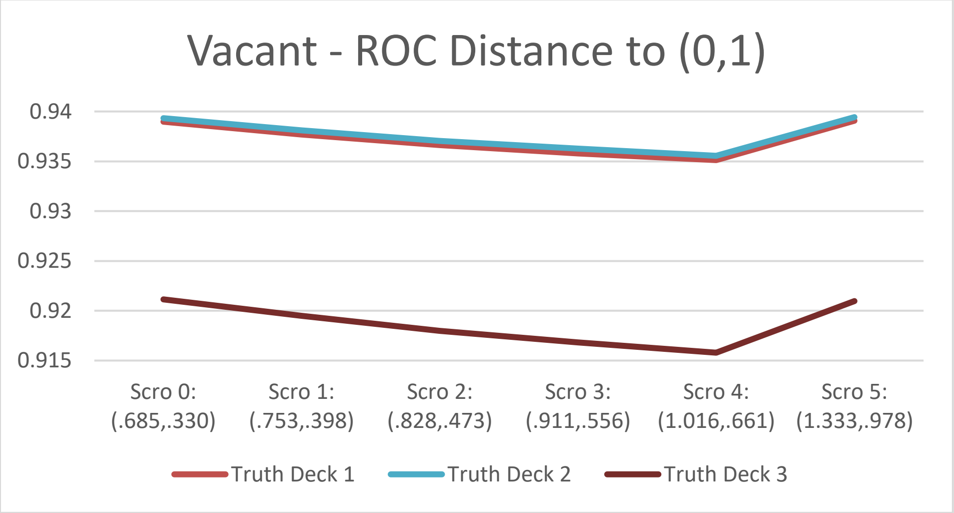
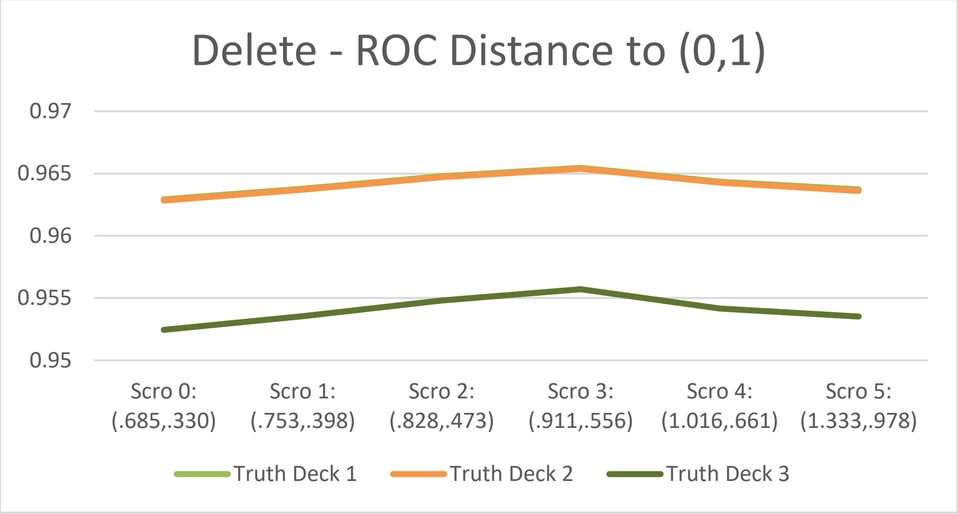


Chart 3: Delete Distance Plots Associated with Perfect Classification Strategy



6. Discussion

Aggregated together, the results indicate that no optimal scenario exists since different metrics indicate different scenarios as the winner. Hence, the choice of the proper use of AR before count imputation may be determined by what is important – e.g. distributional accuracy vs. individual accuracy. One clear lesson is that using no AR or using maximum AR do not provide the best results. Given the caveats above, distances thresholds reflecting the 10% through 20% use of AR appear to be the most promising for developing a cutoff.

This research presents tradeoff options in terms of percentage of AR use scenarios among the unresolved cases. In practice, occupied and vacant/delete distance thresholds need to be specified. If the concern is getting the proper distribution of occupied cases, using between 15% and 20% of AR appears to be the most promising. Translated into an occupied distance threshold, this is somewhere between the 0.9 and 1.0 range. With respect to a vacant/delete distance threshold, using between 15% and 20% of AR maps to a range between 0.55 and 0.66. However, due to the undercount of vacant units, it may be reasonable to extend vacant distance threshold to a range between 0.7 and 0.8.

Note that the final distance thresholds are subject to the unresolved universe that feeds into this analysis. For example, if all unresolved addresses lack AR data, there is no tradeoff that can occur. That is, all records must be imputed. Similarly, a glut of unresolved cases with AR data may result in a higher percentage of AR use.

7. References

Keller, A. and Fox, T. (2012). “2010 Census Coverage Measurement Estimation Report: Components of Census Coverage Results for the Household Population in the United States,” DSSD 2010 Census Coverage Measurement Memorandum Series #2010-G-04.

Keller, A., Mule, V.T., Morris, D.S., Konicki, S. (2018). “A Distance Metric for Modeling the Quality of Administrative Records for Use in the 2020 U.S. Census, Journal of Official Statistics, 34(3), 599-624. DOI: http://dx.doi.org/10.2478/JOS-2018-0029

Metz, Charles E. 1978. “Basic Principles of ROC Analysis.” Seminars in Nuclear Medicine 8:283–98.

United States Census Bureau (2018a). 2020 Census Operational Plan: Version 4.0. Washington DC: Census Bureau. Available at: http://www2.census.gov/programs-surveys/decennial/2020/program-management/planning-docs/2020-oper-plan4.pdf (accessed August 2019).

Williams, T. (2005). “2010 Count Imputation Research – Methodology for Developing the Truth Deck,” DSSD 2006 Census Test Memorandum Series #J2-03.