Census 2010 Count Imputation: Collapsing Strategies Using Mean Squared Error

Andrew Keller

U.S. Census Bureau
4600 Silver Hill Rd.
Washington, DC 20233
andrew.d.keller@census.gov

The views expressed on statistical, methodological, technical, or operational issues are those of the author and not necessarily those of the U.S. Census Bureau.

Abstract
For the 2010 census, the count imputation (CI) procedure filled in missing household status and size for the small proportion of addresses (less than one half percent) where this information was unknown. The CI model partitioned records with complete information defined by household characteristics and geographic area into several cells. We incorporated household characteristics (structure type, enumerator type, nearest neighbor household type) and neighborhood geography (tract) as part of our cell definitions to ensure the unknown addresses were imputed from a distribution of complete addresses with similar characteristics. To account for the varying degrees of information from the unresolved addresses, we constructed three imputation types.

The similarity of housing unit composition achieved by defining imputation cells by household characteristics and geography sometimes involved a tradeoff of creating cells with sparse counts of complete addresses. The essential goal of this research was to use research data from the 2000 Census to aid in developing collapsing procedures for cells with sparse counts. To measure the effectiveness of possible collapsing procedures, we used a mean squared error (MSE) approach. The MSE approach incorporated a composite estimator that accounted for the three types of imputation and estimated bias and variance. To assess different collapsing procedures, we performed simulations for assigning household status and size to missing data cases using two approaches for creating pseudo-missing data. This work outlines potential methodologies for collapsing sparse cells if more complete data addresses are necessary. In addition, it considers possible sizes for the maximum number of complete data addresses that constitute a sparse cell.

Introduction
Count imputation (CI) fills in missing household status and size for addresses where this information is unknown. In 2000, 679,381 (0.55%) of 122,534,761 addresses underwent count imputation. Following Census 2000, research identified an alternative scheme to nearest-neighbor hot deck used in the 2010 Census. Angueira (2008) documents the decision to implement the new methodology identified from the research. Kilmer (2008) provides detailed documentation of the research results.

The new model partitions records defined by household characteristics and geographic area into several cells. These cells formed by these stratifying variables are used to build distributions from complete data addresses (addresses with known household status and size). The distributions are then used to impute status and size for missing data addresses where this information is unknown. To implement this new methodology, we undertook research to better understand how to form cells to minimize error in the imputation process. The essential goal of this research is to compare collapsing procedures. To measure the effectiveness of possible collapsing procedures, we use a mean squared error (MSE) approach. In this paper, we do the following:

- Confirm that, when imputing population size for missing data addresses, using cells with more stratifying variables (and fewer complete data addresses) incurs less error than using cells with fewer stratifying variables (and more complete data addresses)
- Develop a methodology to collapse cells if more complete data addresses are necessary
- Compare possible thresholds for the number of complete data addresses constituting a sparse cell

- Calculate MSE for various combinations of collapsed cells and thresholds

Background

Count imputation methodology is applied to all non-group quarters addresses. Imputed addresses are referred to as CI addresses while addresses not subject to count imputation are non-CI addresses.

Classification Variables

The count imputation model implemented in 2010 assigns structure type, enumerator type, and nearest-neighbor household type variables to CI and non-CI addresses. Household type is only assigned to non-CI addresses since the household composition of CI addresses is unknown. The values for these variables are defined as follows:

Table 1: Definition of Structure Type, Enumerator Type, Household Type, and Nearest-neighbor Household Type

Variable	Possible Values	Comments
Structure Type	0	Single-Unit Structure
	1	Multi-Unit Structure
Enumerator Type[1]	0	Data was captured via personal or phone enumeration, such as Coverage Follow-up Telephone operation, Nonresponse Follow-up operation, Vacant/Delete Check operation.
	1	Data was captured via mail back from respondents, such as an initial mailing, a replacement mailing, or an update/leave operation.
Household Type	1	Household type is a Vacant Address.
	2	Household type is a Married Couple Family Household.
	3	Household type is an Other Family Household.
	4	Household type is a non-Family Household with the Householder Living Alone.
	5	Household type is a non-Family Household with the Householder Not Living Alone.
	6	Household type is an Occupied with Unknown Household Type.
	7	Household type is a Delete or Non-existent Address.
Nearest-neighbor Household Type	1	Nearest Neighbor is a Vacant Address.
	2	Nearest Neighbor is a Married Couple Family Household.
	3	Nearest Neighbor is an Other Family Household.
	4	Nearest Neighbor is a non-Family Household with the Householder Living Alone.
	5	Nearest Neighbor is a non-Family Household with the Householder Not Living Alone.
	6	Nearest Neighbor is an Occupied with Unknown Household Type.
	7	Nearest Neighbor is a Delete or Non-existent Address.

[1] Enumerator Type refers to the last operation in which data about the address was collected.

In addition, both CI and non-CI addresses are geographically mapped to a state, Local Census Office (LCO), county, and collection tract. Within each LCO, the combination of county and collection tract is unique. For the purposes of this work, we refer to the concatenation of county and collection tract as a "tract". State, LCO, tract, structure type, enumerator type, and nearest-neighbor household type are the stratifying variables by which we form a cell. We refer to the state/LCO/tract/structure type/enumerator type/nearest-neighbor household type combination as a tract-cell. Every address (regardless of whether it is missing or complete) is classified into a particular tract-cell. The tract-cell emphasizes the geographic proximity of the CI and non-CI addresses.

Three types of CI addresses exist: status imputation addresses, occupancy imputation addresses, and household size imputation addresses. Status imputations result when the status and size of the CI address are unknown. We do not know if the address represents a valid, livable housing unit. As a result, in order to impute for status imputation addresses, all non-CI addresses (occupied, vacant, and delete) form the applicable distribution. Occupancy imputations result when the CI address is a housing unit (HU), but it is unknown whether the HU is occupied or vacant. As a result, in order to impute for occupancy imputation addresses, occupied and vacant non-CI addresses form the applicable distribution. Note that a vacant non-CI address has a household size of 0 while an occupied non-CI address implies that we know the household size is one or more persons. Household size imputations result when the address is an occupied HU, but household size is unknown. As a result, in order to impute for household size imputation addresses, all occupied non-CI addresses form the applicable distribution. It should be noted that the distributions also use some CI addresses as part of their formation. See Griffin (2007) for the derivation of the distributions for each imputation category. See Pritts (2010) for examples of these calculations.

Methodology

When a CI address exists within the tract-cell, we must first identify what constitutes a sufficient number of non-CI (donor) addresses from the tract-cell distribution for imputation. We call that sufficient number of donor addresses the threshold. In this research, we perform the analysis using various threshold values. If the number of sufficient donor addresses falls below the threshold, we use a distribution from the merged-cell to impute. We form the merged-cell by collapsing over a stratifying variable.

The following example explains how we form the merged-cell. Suppose we have the following two tract-cells with different values of nearest-neighbor household type (NNHT), but the same values for state, LCO, tract, structure type, and enumerator type. In this example, NNHT is the stratifying variable with tract-cell A having NNHT=1 and tract-cell B having NNHT=2.

Tract-Cell A / NNHT=1	Tract-Cell B / NNHT=2
Insufficient Number of Donor Addresses	Sufficient Number of Donor Addresses

Note that an insufficient number of donor addresses exists in tract-cell A. For this example, we collapse over NNHT and use the distribution of donor addresses from tract-cell A <u>and</u> tract-cell B to impute for tract-cell A's CI addresses. However, since a sufficient number of donor addresses exists in tract-cell B, we use the distribution of donor addresses <u>solely</u> from tract-cell B to impute for its CI addresses. There are two items worth noting. First, while we use the distribution of donor addresses from both tract-cells to impute for the CI addresses in tract-cell A, we do not use the distribution of donor addresses in both tract-cells to impute for the CI addresses in tract-cell B. This is different than some imputation schemes that collapse both tract-cells together provided that one cell has an insufficient number of donor addresses. Second, in addition to testing various threshold values, another aspect of this research is geared towards identifying a proper stratifying variable. For this research, we try two stratifying variables to form the merged-cell.

To have a metric with which to answer our questions, we use a MSE approach. MSE is the sum of the squared bias and the variance. In this research, the estimated squared bias quantifies the difference between the household size of non-CI addresses used as donors and CI addresses. The variance quantifies the variability of household size among non-CI addresses. We compare MSE for the tract-cell against the MSE for the merged-cell. If the MSE of the tract-cell is less than the MSE of the merged-cell, the imputation should be performed using the distribution of non-CI addresses within the tract-cell. Conversely, if the MSE of the tract-cell is larger than the MSE of the merged-cell, the imputation should be performed using the distribution of non-CI addresses within the merged-cell.

Rather than perform separate analysis for each type of CI address, the MSE approach uses a composite estimator to calculate bias and variance while accounting for the three types of imputation. Attachment A contains the derivations of MSE for the tract-cell and merged-cell. Attachment B provides a numerical example to demonstrate the computations used in this research.

<u>Merged-Cell Possibilities</u>

We consider two merged-cell possibilities for this work. We believe it is important to maintain geographic proximity between CI addresses and the distribution of donor addresses. Consequently, the first merged-cell collapses over nearest-neighbor household type when forming the merged-cell. As a result, the merged-cell distribution is created from all donor addresses sharing the same state/LCO/tract/structure type/enumerator type. We call this the **Nearest-Neighbor Merged-Cell**. Meanwhile, the tract-cell is created from all donor addresses sharing the same state/LCO/tract/structure type/enumerator type/nearest-neighbor household type.

The second merged-cell deemphasizes geographic proximity between CI addresses and the distribution of donor addresses. It does this by collapsing over tract when forming the merged-cell. As a result, the merged-cell distribution is created from all donor addresses sharing the same state/LCO/structure type/enumerator type/nearest-neighbor household type. We call this the **Tract Merged-Cell**. In short, we believe this research will show that using the **Nearest-Neighbor Merged-Cell** (and maintaining geographic proximity) results in less error than using the **Tract Merged-Cell**.

<u>Designating Household Sizes for CI Addresses</u>

As can be seen from Attachment A, in order to do this MSE analysis we had to incorporate calculations including CI addresses. By definition, CI addresses have unknown household size. To compensate for this, we had to designate

some donor (or non-CI) addresses as CI addresses. These designated donor addresses are called pseudo-CI households. We completed this designation using two approaches to simulate missingness. These approaches resulted in two datasets.

For the first dataset, we used a truth deck approach. To research different count imputation methodologies for Census 2010, a truth deck was created from Census 2000 data. The truth deck takes a small percentage (~1%) of donor households and changes them to pseudo-CI households. The pseudo-CI households are not randomly designated. Rather, the pseudo-CI housing units are those which have characteristics commonly associated with having household status or size or both missing. Testing of count imputation methodology is then completed using pseudo-CI households to represent CI addresses as if we know the true size of the CI household. We call this the **Truth Deck CI Dataset**. See Williams (2005) for more information on how the donor addresses are designated as pseudo-CI addresses.

For the second dataset, we used an approach designating donor addresses processed last as the pseudo-CI addresses. The rationale behind this decision was that the latest donor addresses could have been CI addresses if the Census Bureau had closed its operations slightly earlier. We believe these cases may have composition somewhat similar to the actual CI addresses. We call this the **Late CI Dataset**.

With the 2000 data, most non-CI addresses have a processing date designating when the household was entered into the census processing system. Within a tract-cell, we assigned the latest non-CI addresses as the pseudo-CI addresses provided that the non-CI address' unit status was applicable for the CI address. For example:

1) For status imputation addresses, all non-CI addresses were eligible to be designated as pseudo-CI addresses.
2) For occupancy imputation addresses, both occupied units with household size information and vacant non-CI addresses were eligible to be designated as pseudo-CI addresses.
3) For household size imputation addresses, only occupied non-CI addresses with household size information were eligible to be designated as pseudo-CI addresses.

We geared this research towards understanding when to collapse cells with CI addresses to ensure a sufficient number of non-CI addresses to define a distribution. Consequently, we restricted this analysis only to tract-cells with both non-CI and CI addresses. With respect to the number of non-CI and CI addresses, three tract-cell combinations exist: tract-cells with more non-CI addresses than CI addresses, tract-cells with fewer non-CI addresses than CI addresses, and tract-cells with an equal number of non-CI and CI addresses. Because the imputation rate is only 0.55%, the first type is far more prevalent than the latter two.

The Truth Deck CI Dataset is straightforward in its designation of pseudo-CI addresses from donor addresses. The example in Attachment B shows the relevant calculations based on formulas from Attachment A for the Truth Deck CI Dataset. However, the Late CI Dataset requires intricate data manipulation to designate pseudo-CI addresses from donor addresses. To illustrate this, Attachment C presents examples of the three types of tract-cells discussed in the previous paragraph by showing how the donor and pseudo-CI addresses were designated for the Late CI Dataset. Note that we always retain the earliest non-CI address as a non-CI address, meaning not to assign this non-CI address as a pseudo-CI address even if the number of non-CI addresses is less than the number of CI addresses in the tract-cell. These examples also demonstrate how the relevant calculations from Attachment A were completed. In addition, exclusive to the Late CI Dataset, cells with one applicable non-CI address and CI addresses were excluded from this analysis. Attachment C provides the rationale for this decision via two examples.

The Truth Deck CI Dataset and Late CI Dataset are implemented from different datasets created from the 2000 Census results. In short, the non-CI addresses and CI addresses are different between the two datasets. Hence, it is not appropriate to compare across datasets. The comparison of interest involves comparing a) between tract-cell and merged-cell approaches within thresholds and b) among thresholds within each dataset.

Results

We use this research to help identify a suitable threshold to understand when to impute from the distribution of cases within the tract-cell as opposed to the distribution of cases within the broader merged-cell. However, there are certain situations where we decide to use the tract- or merged-cell distribution to impute. This course of action depends on the number of donor addresses in the tract-cell.

For example, suppose we have 50 applicable donor addresses within a tract-cell. In this instance, we decide to always impute the CI addresses from the donor addresses within the tract-cell because a sufficient number of donor addresses exist. Alternatively, suppose we have only two applicable donor addresses in the tract-cell. In this instance, we decide not to impute CI addresses from the tract-cell based on the information of only two donor addresses. This research examines various threshold values to better understand outcomes with respect to bias, variance, and mean squared error for the count imputation methodology.

Table 2 and Table 3 show results for the two ways of assigning missingness for pseudo-CI addresses, the **Truth Deck CI Dataset** and the **Late CI Dataset** respectively. Column (A) lists the three thresholds we examined (10, 20, 30). Column (B) shows the applicable number of cells for which the number of donor addresses falls below the given threshold and has at least one CI address. For each Threshold and Applicable Cells pairing, Column C lists the three types of cells for computing the metrics: Tract-cell, Nearest-Neighbor Merged-Cell, and Tract Merged-Cell. Columns (D), (E), and (F) list the average squared bias, average variance, average MSE for each possible combination. Column (G) shows the proportion of times where the tract-cell has a lower MSE than the merged-cell. See Attachment D for the computation of the average squared bias and variance for the tract-cell, average squared bias and variance for the merged-cell, and percentage of cells in which the MSE for the tract-cell is less than the MSE for its corresponding merged-cell.

For the **Truth Deck CI Dataset** and Threshold=10/12,555 Applicable Cells pairing, we compare the relevant statistics in columns (D), (E), and (F) describing the tract-cell against the relevant statistics describing the two merged-cells. For those 12,555 tract-cells, the average squared bias is 3.874, the average variance is 1.743, and the average MSE is 5.617. For those same tract-cells, when merging over nearest-neighbor type, the corresponding nearest-neighbor merged cells have an average squared bias of 2.926, an average variance of 2.501, and an average MSE of 5.427. For this combination, 58.15% of the 12,555 applicable tract-cells have an MSE less than that of its corresponding nearest-neighbor merged-cell.

This result seems counterintuitive. Since the average MSE across all 12,555 tract-cells is greater than the average MSE across all corresponding nearest-neighbor merged-cells, one would expect that less than 50% of the tract-cells would have a lower MSE than its corresponding merged-cell. This does not occur. In short, a few tract-cells have large squared bias and variance values. These few tract-cells inflate the average squared bias and variance even though the general trend across the 12,555 cells is that the MSE of the tract-cell is less than that of its corresponding nearest-neighbor merged-cell.

By contrast, when merging over tract, the corresponding tract-merged cells have an average squared bias average squared bias of 3.025, an average variance of 2.776, and an average MSE of 5.801. For this combination, 63.26% of the 12,555 applicable tract-cells have an MSE less than that of its corresponding merged-cell. With the tract merged-cell, this result seems more sensible. That is, since the average MSE of the tract-cell is less than the average MSE of the tract-merged-cell, one would expect that more than 50% of the tract-cells would have a lower MSE than its corresponding tract merged-cell. This occurs.

Next, for the **Truth Deck CI Dataset** and Threshold=20/29,553 Applicable Cells combination, we compare the relevant statistics in columns (D), (E), and (F) describing the tract-cell against the relevant statistics describing the two merged-cells. For those 29,553 tract-cells, the average squared bias is 3.518, the average variance is 2.266, and the average MSE is 5.784. For those same tract-cells, when merging over nearest-neighbor type, the corresponding merged cells have an average squared bias of 3.065, an average variance of 2.712, and an average MSE of 5.777. For this combination, 56.83% of the 29,553 applicable tract-cells have an MSE less than that of its corresponding nearest-neighbor merged-cell. By contrast, when merging over tract, their corresponding tract merged-cells have an average squared bias of 3.140, an average variance of 2.912, and an average MSE of 6.052. For this combination, 62.44% of the 29,553 applicable tract-cells have an MSE less than that of its corresponding merged-cell.

Table 2: Comparing MSE Results from Truth Deck CI Dataset

Threshold (A)	Applicable Number of Tract-Cells (B)	Cell Type (C)	Average Squared Bias (D)	Average Variance (E)	Average MSE (F)	% Cells where MSE of Tract-cell < MSE of Merged-cell for Applicable Cells (G)
10	12,555	Tract-Cell	3.874	1.743	5.617	---
		Nearest-Neighbor Merged-Cell	2.926	2.501	5.427	58.15%
		Tract Merged-Cell	3.025	2.776	5.801	63.26%
20	29,553	Tract-Cell	3.518	2.266	5.784	---
		Nearest-Neighbor Merged-Cell	3.065	2.712	5.777	56.83%
		Tract Merged-Cell	3.140	2.912	6.052	62.44%
30	47,191	Tract-Cell	3.390	2.495	5.885	---
		Nearest-Neighbor Merged-Cell	3.100	2.813	5.913	56.26%
		Tract Merged-Cell	3.246	3.002	6.248	62.01%

Table 3: Comparing MSE Results from Late CI Dataset

Threshold (A)	Applicable Number of Tract-Cells (B)	Cell Type (C)	Average Squared Bias (D)	Average Variance (E)	Average MSE (F)	% Cells where MSE of Tract-cell < MSE of Merged-cell for Applicable Cells (G)
10	20,749	Tract-Cell	109.414	2.198	111.612	---
		Nearest-Neighbor Merged-Cell	51.874	4.766	56.640	56.03%
		Tract Merged-Cell	33.321	6.793	40.114	61.11%
20	43,069	Tract-Cell	76.131	2.924	79.055	---
		Nearest-Neighbor Merged-Cell	37.568	4.535	42.104	55.50%
		Tract Merged-Cell	25.577	5.809	31.386	61.43%
30	63,180	Tract-Cell	58.050	3.280	61.330	---
		Nearest-Neighbor Merged-Cell	29.933	4.525	34.458	55.11%
		Tract Merged-Cell	22.428	5.555	27.983	61.15%

A few results stand out from Tables 2 and 3. First, note that we have percentages greater than 50% in column (G). So, based on the defined methodology, we know that for all addresses (regardless merged-cell type or the dataset by which we designated household sizes for CI addresses), the MSE of the tract-cell is less than the MSE of its corresponding merged-cell more than half the time. In general, this result confirms that, when using either of the two types of merged-cells, we incur more error if we were to impute using the merged-cell. Furthermore, when comparing within each dataset, in column (G) we generally see that the percentage of cells where the MSE of the tract-cell is less than the MSE of the merged-cell decreases as the threshold increases. This shows that, as the threshold increases, more tract-cells will have larger MSE values than their corresponding merged cells.

Second, when comparing the **Nearest-Neighbor Merged-Cell** type versus **Tract Merged-Cell** type for both datasets, the higher percentages seen in the **Tract Merged-Cell** type rows imply that the MSE is generally larger when we define the merged-cell by collapsing over tracts (**Tract Merged-Cell** type) rather than collapsing over nearest-neighbor household type (**Nearest-Neighbor Merged-Cell** type). This result suggests that it would be better to define the merged-cell by collapsing over nearest-neighbor household type. Or, more generally, it would be better to define the merged-cell by maintaining geographical proximity (in this case, tract).

The final objective of this research was to develop the threshold for the minimum number of donor addresses when using the tract-cell to impute. From Tables 2 and 3, we generally observe that the tract-cells have smaller MSE values than their corresponding merged-cells at lower threshold levels. Consequently, it would be reasonable to require the lowest threshold possible in order to impute using the tract-cell. The extreme case would be requiring only one applicable non-CI address as a donor for imputing all CI addresses.

Keller (2010) shows that 17% of CI addresses are grouped together in clusters of length 11 or greater. Because CI addresses are often clustered together, it is debatable whether requiring only one donor address in the tract-cell to impute for all CI addresses would be prudent. If this were the case, the cluster of CI addresses would be imputed with the same household size. For example, suppose we set the requirement to be one donor address. Additionally, suppose we must impute in a tract-cell with ten CI addresses and one donor address with a household size of 9. For this hypothetical tract-cell, we would impute each of the ten CI addresses with a household size of 9. To avoid scenarios like above, we looked at 10 as the minimum threshold for our research.

For the Truth Deck CI Dataset, as the threshold increases, the MSE increases within the tract-cell. Conversely, for the Late CI Dataset, as the threshold increases, the MSE decreases within the tract-cell. However, for both CI datasets, as the threshold increases, the squared bias decreases and the variance increases within the tract-cell. Therefore, if relying upon the MSE as a metric to decide upon the threshold, the conclusion is vague because of the inconsistency depending on the dataset used.

Conclusions

The purpose of this research was threefold. First, in the column (G) of Table 2 and Table 3 we confirmed that tract-cells incur less error than merged-cells when imputing household size for CI addresses. This was to be expected since tract-cells are believed to be more homogenous than merged-cells with respect to household size. Second, we obtained ideas on how to best collapse tract-cells to form merged-cells. We learned that it is best to define the merged-cell by collapsing over nearest-neighbor household type rather than by collapsing over tract. Last, we sought to identify a threshold for the number of donor addresses in a tract-cell in order to impute from the tract-cell. The conclusions stemming from these results are less clear. With respect to MSE, we produce different conclusions on where to set the threshold. This depends on the CI Dataset employed.

With regard to future work, two items stand out. First, we used MSE as our metric of evaluation for this research. It is possible to use other metrics. For example, instead of doing the analysis using squared loss over the nation, we can analyze our results by calculating absolute loss with respect to state population counts. Second, instead of organizing the latest non-CI addresses by capture date and designating them as pseudo-CI addresses, we could have designated the pseudo-CI addresses by type of operation. That is, we could have assigned all the pseudo-CI addresses if they were part of a certain census operation like Non-Response Followup.

References

Angueira, T. (2008). "Choice of Count Imputation Methodology for the 2010 Census," DSSD 2010 DECENNIAL CENSUS PROGRAM DECISION MEMORANDUM SERIES No. 22.

Griffin, R. (2007). "2010 Count Imputation Research - Modified Imputation Methodology Using Spatial Modeling Results," DSSD 2006 CENSUS TEST MEMORANDUM SERIES J2–13.

Keller, A. (2010). "Census 2010 Count Imputation: Issues Concerning Imputation Levels and Clustering," DSSD 2010 DECENNIAL CENSUS MEMORANDUM SERIES J–07.

Kilmer, A. (2008). "2010 Census Count Imputation Research - Results and Analysis of Count Imputation Methodologies for All States," DSSD 2006 CENSUS TEST MEMORANDUM SERIES J2 - 21.

Pritts, M. (2010). "Census 2010: Overview of Count Imputation," DSSD 2010 DECENNIAL CENSUS MEMORANDUM SERIES J–08.

Williams, T.R. (2005). "2010 Count Imputation Research – Methodology for Developing the Truth Deck," DSSD 2006 CENSUS TEST MEMORANDUM SERIES J2–03.

Attachment A

The following methodology was developed by Patrick Cantwell of the Census Bureau.

To begin, let *k* refer to the tract-cell and *M* refer to the merged-cell. Therefore, we merge if:

$$MSE_k > MSE_M$$

$$\Leftrightarrow Bias_k^2 + Variance_k > Bias_M^2 + Variance_M \qquad (1)$$

Now, rewriting the bias, let $y_{i,k}$ refer to the total number of person records in a housing unit *i* in tract-cell *k*, $\hat{y}_{i,k}$ refer to the estimate of total number of person records in a housing unit *i* in tract-cell *k*, u_k the set of non-CI housing units in tract-cell *k*, and Y_k the true unknown total number of person records in housing units in tract-cell *k*. We condition on u_k because it is a known total for each tract-cell. Then,

$$\begin{aligned} Bias_k &= E(\hat{y}_k \mid u_k) - Y_k \\ &= E\left[\left(\sum_{i\in non-CI} y_{i,k} + \sum_{i\in CI} \hat{y}_{i,k}\right) \mid u_k\right] - \left[\sum_{i\in non-CI} y_{i,k} + \sum_{i\in CI} y_{i,k}\right] \\ &= \sum_{i\in non-CI} y_{i,k} + E\left[\sum_{i\in CI} \hat{y}_{i,k} \mid u_k\right] - \sum_{i\in non-CI} y_{i,k} - \sum_{i\in CI} y_{i,k} = E\left[\sum_{i\in CI} \hat{y}_{i,k} \mid u_k\right] - \sum_{i\in CI} y_{i,k} \\ &= \sum_{i\in CI} E\left[\hat{y}_{i,k} \mid u_k\right] - \sum_{i\in CI} y_{i,k} \qquad (2) \end{aligned}$$

It's important to note that we have three separate types of count imputation addresses: status imputations, occupancy imputations, and household size imputations. As a result, we must compensate for all three types when we make the decision to merge cells. So continuing from (2),

$$\begin{aligned} Bias_k &= \sum_{i\in CI} E\left[\hat{y}_{i,k} \mid u_k\right] - \sum_{i\in CI} y_{i,k} \\ &= \sum_{i\in SI} E\left[\hat{y}_{i,k}^{(1)} \mid u_k^{(1)}\right] - \sum_{i\in SI} y_{i,k}^{(1)} + \sum_{i\in OccI} E\left[\hat{y}_{i,k}^{(2)} \mid u_k^{(2)}\right] - \sum_{i\in OccI} y_{i,k}^{(2)} + \sum_{i\in HHSI} E\left[\hat{y}_{i,k}^{(3)} \mid u_k^{(3)}\right] - \sum_{i\in HHSI} y_{i,k}^{(3)} \qquad (3) \end{aligned}$$

where

$i \in SI, i \in OccI \in k, i \in HHSI : i$ is a status, occupancy, or household size pseudo-CI address in cell *k*

$\hat{y}_{i,k}^{(1)}$: the estimate of total number of person records in tract-cell k imputed from status imputation

$u_k^{(1)}$: the set of non-CI housing units in tract-cell *k*

$\hat{y}_{i,k}^{(2)}$: the estimate of total number of person records in tract-cell k imputed from occupancy imputation

$u_k^{(2)}$: the set of non-CI, non-delete housing units in tract-cell $k-$ delete cases are excluded since occupancy imputation implies that we know this address is a housing unit but we are unsure whether it is occupied or vacant

$\hat{y}_{i,k}^{(3)}$: the estimate of total number of person records in tract-cell k imputed from household size imputation

$u_k^{(3)}$: the set of non-CI, occupied housing units with household size information in tract-cell $k-$ delete and vacant cases are excluded since household size imputation implies that we know this address is an occupied housing unit but we are unsure of its household size

Looking at the first term in (3), we know that the total <u>*expected*</u> number of person records added by status imputation in tract-cell *k* is the average of all non-CI addresses multiplied by the number of addresses requiring status imputation. That is,

$\sum_{i\in SI} E\left[\hat{y}_{i,k}^{(1)} \mid u_k^{(1)}\right] = SI_k \bar{y}_{non-CI,k}^{(1)}$ where SI_k refers to the number of status imputation addresses in cell *k* and $\bar{y}_{non-CI,k}^{(1)}$ refers to the average household size of all non-CI addresses in cell *k*.

Now, looking at the second term in (3), we know that the total number of person records added by status imputation in tract-cell *k* is the average household size of all status imputation addresses multiplied by the number of addresses requiring status imputation. That is,

$\sum_{i \in SI} y_{i,k}^{(1)} = SI_k \bar{y}_{SI,k}$ where $\bar{y}_{SI,k}$ refers to the average household size of all status imputation addresses in cell *k*.

Note that since the household size of any imputation address is unknown, for this research we use pseudo-CI addresses to form the basis for any calculations dealing with CIs. So, to calculate $\bar{y}_{SI,k}$: $\bar{y}_{SI,k} = \frac{1}{SI_k} \sum_{i \in SI} y_{i,k}$.

Similar calculations are completed for:

a) $\sum_{i \in OccI} E\left[\hat{y}_{i,k}^{(2)} \mid u_k^{(2)}\right] - \sum_{i \in OccI} y_{i,k}^{(2)} = OccI_k \bar{y}_{non-CI,k}^{(2)} - OccI_k \bar{y}_{OccI,k}$

where

$OccI_k$: the number of occupancy imputation addresses in cell *k*

$\bar{y}_{non-CI,k}^{(2)}$: the average household size of all non-delete, non-CI addresses in cell *k*.

$\bar{y}_{OccI,k}$: the average household size of all occupancy imputation addresses in cell *k*.

b) $\sum_{i \in HHSI} E\left[\hat{y}_{i,k}^{(3)} \mid u_k^{(3)}\right] - \sum_{i \in HHSI} y_{i,k}^{(3)} = HHSI_k \bar{y}_{non-CI,k}^{(3)} - HHSI_k \bar{y}_{HHSI,k}$

where

$HHSI_k$: the number of household size imputation addresses in cell *k*

$\bar{y}_{non-CI,k}^{(3)}$: the average household size of all occupied non-CI addresses in cell *k*.

$\bar{y}_{HHSI,k}$: the average household size of all household size imputation addresses in cell *k*.

Hence, we can rewrite (3) as:

$$\begin{aligned} Bias_k &= \sum_{i \in SI} E\left[\hat{y}_{i,k}^{(1)} \mid u_k^{(1)}\right] - \sum_{i \in SI} y_{i,k}^{(1)} + \sum_{i \in OccI} E\left[\hat{y}_{i,k}^{(2)} \mid u_k^{(2)}\right] - \sum_{i \in OccI} y_{i,k}^{(2)} + \sum_{i \in HHSI} E\left[\hat{y}_{i,k}^{(3)} \mid u_k^{(3)}\right] - \sum_{i \in HHSI} y_{i,k}^{(3)} \\ &= \left[SI_k \bar{y}_{non-CI,k}^{(1)} - SI_k \bar{y}_{SI,k}\right] + \left[OccI_k \bar{y}_{non-CI,k}^{(2)} - OccI_k \bar{y}_{OccI,k}\right] + \left[HHSI_k \bar{y}_{non-CI,k}^{(3)} - HHSI_k \bar{y}_{HHSI,k}\right] \end{aligned} \tag{4}$$

Now, the total number of CI addresses is the sum of Status, Occupancy, and Household Size imputation addresses. That is, $CI_k = SI_k + OccI_k + HHSI_k$. So, we can rewrite (4) as:

$$\begin{aligned} Bias_k &= \left[SI_k \bar{y}_{non-CI,k}^{(1)} - SI_k \bar{y}_{SI,k}\right] + \left[OccI_k \bar{y}_{non-CI,k}^{(2)} - OccI_k \bar{y}_{OccI,k}\right] + \left[HHSI_k \bar{y}_{non-CI,k}^{(3)} - HHSI_k \bar{y}_{HHSI,k}\right] \\ &= SI_k \left[\bar{y}_{non-CI,k}^{(1)} - \bar{y}_{SI,k}\right] + OccI_k \left[\bar{y}_{non-CI,k}^{(2)} - \bar{y}_{OccI,k}\right] + HHSI_k \left[\bar{y}_{non-CI,k}^{(3)} - \bar{y}_{HHSI,k}\right] \\ &= CI_k \left[\frac{SI_k}{CI_k}\left[\bar{y}_{non-CI,k}^{(1)} - \bar{y}_{SI,k}\right] + \frac{OccI_k}{CI_k}\left[\bar{y}_{non-CI,k}^{(2)} - \bar{y}_{OccI,k}\right] + \frac{HHSI_k}{CI_k}\left[\bar{y}_{non-CI,k}^{(3)} - \bar{y}_{HHSI,k}\right]\right] \end{aligned} \tag{5}$$

Similarly, for the merged-cell we can write:

$$Bias_M = CI_k \left[\frac{SI_k}{CI_k}\left[\bar{y}_{non-CI,M}^{(1)} - \bar{y}_{SI,k}\right] + \frac{OccI_k}{CI_k}\left[\bar{y}_{non-CI,M}^{(2)} - \bar{y}_{OccI,k}\right] + \frac{HHSI_k}{CI_k}\left[\bar{y}_{non-CI,M}^{(3)} - \bar{y}_{HHSI,k}\right]\right]$$

Where $\bar{y}_{non-CI,M}^{(1)}, \bar{y}_{non-CI,M}^{(2)}, \bar{y}_{non-CI,M}^{(3)}$ are defined analogously over the merged-cell.

Looking at the variance term in (1), we assume that the CI addresses are independent of one another. This is because we make an independent random draw to impute for each CI address. As a result, we can write:

$$Variance_k = Var\left[\left(\sum_{i \in non-CI} y_{i,k} + \sum_{i \in CI} \hat{y}_{i,k}\right) \mid u_k\right] = Var\left[\sum_{i \in CI} \hat{y}_{i,k} \mid u_k\right] = \sum_{i \in CI} Var\left[\hat{y}_{i,k} \mid u_k\right] \tag{6}$$

Furthermore, we assume the imputed result for each CI address follows as from a simple random sample from the non-CI addresses within the cell. As a result, we can write the variance for each individual CI address as:

$Var(\hat{y}_k \mid u_k) = \left(1 - \frac{1}{n_{non-CI,k}}\right)\frac{S_k^2}{1} = \left(\frac{n_{non-CI,k} - 1}{n_{non-CI,k}}\right)\frac{S_k^2}{1} = S_k^{'2}$ where $S_k^{'2}$ reflects the addresses in the cell considered as a population and is defined as $S_k^{'2} = \frac{1}{n_{non-CI,k}} \sum_{i=1}^{n_{non-CI,k}} \left[y_{non-CI,k,i} - \bar{y}_{non-CI,k}\right]^2$.

Now, rewriting (6) including all three types of imputation:

$$Variance_k = \sum_{i \in CI} Var\left[\hat{y}_{i,k} \mid u_k\right] = \sum_{i \in SI} Var\left[\hat{y}_{i,k}^{(1)} \mid u_k^{(1)}\right] + \sum_{i \in OccI} Var\left[\hat{y}_{i,k}^{(2)} \mid u_k^{(2)}\right] + \sum_{i \in HHSI} Var\left[\hat{y}_{i,k}^{(3)} \mid u_k^{(3)}\right]$$

$$= SI_k S_k^{'2(1)} + OccI_k S_k^{'2(2)} + HHSI_k S_k^{'2(3)} \tag{7}$$

where $S_k^{'2(1)} = \frac{1}{n_{non-CI,k}^{(1)}} \sum_{i=1}^{n_{non-CI,k}^{(1)}} \left[y_{non-CI,k,i}^{(1)} - \bar{y}_{non-CI,k}^{(1)}\right]^2$ and similarly for $S_k^{'2(2)}, S_k^{'2(3)}$

Continuing from (7), we can write:

$$Variance_k = SI_k S_k^{'2(1)} + OccI_k S_k^{'2(2)} + HHSI_k S_k^{'2(3)}$$

$$= CI_k \left[\frac{SI_k}{CI_k} S_k^{'2(1)} + \frac{OccI_k}{CI_k} S_k^{'2(2)} + \frac{HHSI_k}{CI_k} S_k^{'2(3)}\right]$$

Similarly, for the merged-cell we can write:

$$Variance_M = CI_k \left[\frac{SI_k}{CI_k} S_M^{'2(1)} + \frac{OccI_k}{CI_k} S_M^{'2(2)} + \frac{HHSI_k}{CI_k} S_M^{'2(3)}\right]$$

Overall, for the merge rule, we can rewrite:

$$Bias_k^2 + Variance_k > Bias_M^2 + Variance_M$$

$$\Leftrightarrow Bias_M^2 - Bias_k^2 < Variance_k - Variance_M$$

$$\Leftrightarrow \left(CI_k \left[\frac{SI_k}{CI_k}\left[\bar{y}_{non-CI,M}^{(1)} - \bar{y}_{SI,k}\right] + \frac{OccI_k}{CI_k}\left[\bar{y}_{non-CI,M}^{(2)} - \bar{y}_{OccI,k}\right] + \frac{HHSI_k}{CI_k}\left[\bar{y}_{non-CI,M}^{(3)} - \bar{y}_{HHSI,k}\right]\right]\right)^2$$

$$-\left(CI_k \left[\frac{SI_k}{CI_k}\left[\bar{y}_{non-CI,k}^{(1)} - \bar{y}_{SI,k}\right] + \frac{OccI_k}{CI_k}\left[\bar{y}_{non-CI,k}^{(2)} - \bar{y}_{OccI,k}\right] + \frac{HHSI_k}{CI_k}\left[\bar{y}_{non-CI,k}^{(3)} - \bar{y}_{HHSI,k}\right]\right]\right)^2 <$$

$$CI_k \left[\frac{SI_k}{CI_k} S_k^{'2(1)} + \frac{OccI_k}{CI_k} S_k^{'2(2)} + \frac{HHSI_k}{CI_k} S_k^{'2(3)}\right] - CI_k \left[\frac{SI_k}{CI_k} S_M^{'2(1)} + \frac{OccI_k}{CI_k} S_M^{'2(2)} + \frac{HHSI_k}{CI_k} S_M^{'2(3)}\right]$$

$$\Leftrightarrow \left(\begin{array}{l} \left[\frac{SI_k}{CI_k}\left[\bar{y}_{non-CI,M}^{(1)} - \bar{y}_{SI,k}\right] + \frac{OccI_k}{CI_k}\left[\bar{y}_{non-CI,M}^{(2)} - \bar{y}_{OccI,k}\right] + \frac{HHSI_k}{CI_k}\left[\bar{y}_{non-CI,M}^{(3)} - \bar{y}_{HHSI,k}\right]\right]^2 \\ -\left[\frac{SI_k}{CI_k}\left[\bar{y}_{non-CI,k}^{(1)} - \bar{y}_{SI,k}\right] + \frac{OccI_k}{CI_k}\left[\bar{y}_{non-CI,k}^{(2)} - \bar{y}_{OccI,k}\right] + \frac{HHSI_k}{CI_k}\left[\bar{y}_{non-CI,k}^{(3)} - \bar{y}_{HHSI,k}\right]\right]^2 \end{array}\right) <$$

$$\left(\frac{1}{CI_k}\right)\left(\left[\frac{SI_k}{CI_k} S_k^{'2(1)} + \frac{OccI_k}{CI_k} S_k^{'2(2)} + \frac{HHSI_k}{CI_k} S_k^{'2(3)}\right] - \left[\frac{SI_k}{CI_k} S_M^{'2(1)} + \frac{OccI_k}{CI_k} S_M^{'2(2)} + \frac{HHSI_k}{CI_k} S_M^{'2(3)}\right]\right)$$

Attachment B

The following example is provided to illustrate the application of formulas given in Attachment A.

Suppose we have the tract-cell with three CI addresses, two status imputation addresses and one occupancy imputation address. Suppose further that the tract-cell has 16 non-CI addresses with the following distribution of addresses:

Delete	Size=1	Size=2	Size=3	Size=4	Size=5	Size=6	Vacant
1	2	4	3	2	1	1	2

Using the notation from Attachment A then,

$$SI_k = 2, OccI_k = 1, HHSI_k = 0$$

$$\bar{y}^{(1)}_{non-CI,k} = 2.375, \bar{y}^{(2)}_{non-CI,k} = 2.533, \bar{y}^{(3)}_{non-CI,k} = 2.923$$

$$S_k^{'2(1)} = 2.984, S_k^{'2(2)} = 2.782, S_k^{'2(3)} = 2.071$$

Now, suppose that we can designate the household sizes for the CI addresses. We explain the methodology to designate household sizes for the CI addresses on pages 2 and 3 in the main text. Suppose that the two status imputation addresses are size 3 and 4 respectively. Suppose that the occupancy imputation address is size 2. Then,

$$\begin{aligned} Bias_k &= CI_k \left[\frac{SI_k}{CI_k}\left[\bar{y}^{(1)}_{non-CI,k} - \bar{y}_{SI,k}\right] + \frac{OccI_k}{CI_k}\left[\bar{y}^{(2)}_{non-CI,k} - \bar{y}_{OccI,k}\right] + \frac{HHSI_k}{CI_k}\left[\bar{y}^{(3)}_{non-CI,k} - \bar{y}_{HHSI,k}\right] \right] \\ &= 3\left[\frac{2}{3}(2.375-3.5) + \frac{1}{3}(2.533-2)\right] \\ &= -1.7167 \end{aligned}$$

$$\begin{aligned} Variance_k &= CI_k \left[\frac{SI_k}{CI_k} S_k^{'2(1)} + \frac{OccI_k}{CI_k} S_k^{'2(2)} + \frac{HHSI_k}{CI_k} S_k^{'2(3)} \right] \\ &= 3\left[\frac{2}{3}(2.984) + \frac{1}{3}(2.782)\right] \\ &= 8.751 \end{aligned}$$

Similar calculations can be done for $Bias_M^2$ and $Variance_M$ considering the merged-cell data.

Attachment C

Attachment C shows examples of data manipulation for cells that were analyzed via the Late CI Method. Example 1 shows a tract-cell with more non-CI addresses than CI addresses. Example 2 shows a tract-cell with fewer non-CI addresses than CI addresses. Example 3 shows a tract-cell with an equal number of non-CI and CI addresses. Example 4 and 5 show tract-cells where only one non-CI address is applicable and explains why these cells are excluded from the analysis.

Example 1 – Tract-Cell with more non-CI addresses than CI addresses

This is an example tract-cell with 15 non-CI addresses and 5 CI addresses (2 status imputation addresses, 2 occupancy imputation addresses, and 1 household size imputation address).

					Selected As Pseudo-CI Address by Type		
ID	CI Address	Household (HH) Type	HH Size	Return Date	Status	Occupancy	HH Size
1	Status Imputation						
2	Status Imputation						
3	Household Size Imputation						
4	non-CI address	Occupied	2	22-Mar			
5	non-CI address	Occupied	1	23-Mar			
6	non-CI address	Occupied	1	24-Mar			
7	non-CI address	Delete	0	25-Mar			
8	non-CI address	Vacant	0	26-Mar			
9	non-CI address	Vacant	0	27-Mar			
10	non-CI address	Occupied	2	28-Mar			
11	non-CI address	Occupied	4	29-Mar			
12	Occupancy Imputation						
13	non-CI address	Vacant	0	30-Mar			
14	non-CI address	Occupied	3	31-Mar			
15	non-CI address	Occupied	4	1-Apr			
16	Occupancy Imputation						
17	non-CI address	Vacant	0	10-Apr			
18	non-CI address	Occupied	2	11-Apr		Pseudo-CI	
19	non-CI address	Occupied	6	12-Apr	Pseudo-CI	Pseudo-CI	Pseudo-CI
20	non-CI address	Delete	0	13-Apr	Pseudo-CI		

Using the notation from Attachment A, we give the values for all relevant terms in the tract-cell with an explanation:

Table C.1 – Values for Relevant Terms in Example 1

Term	Value	Explanation
SI_k	2	ID #1 and ID #2 are status imputation addresses.
$OccI_k$	2	ID #12 and ID #16 are occupancy imputation addresses.
$HHSI_k$	1	ID #3 is a household size imputation address.
$\bar{y}^{(1)}_{non-CI,k}$	1.462	This is the average household size of ID #4 through #11, ID #13 through #15, ID #17 through #18. Note that ID #19 through #20 are excluded from the calculation because they are part of the calculation of $\bar{y}_{SI,k}$.
$\bar{y}_{SI,k}$	3	ID #19 and #20 are the two latest non-CI addresses. Hence, they are part of this calculation.
$\bar{y}^{(2)}_{non-CI,k}$	1.545	This is the average household size of ID #4 through #6, ID #8 through #11, ID #13 through #15, and ID #17. Note that ID #18 through #19 are excluded from the calculation because they are part of the calculation of $\bar{y}_{OccI,k}$. ID#7 is excluded from this calculation because it is a delete address and not applicable.
$\bar{y}_{OccI,k}$	4	ID #18 and #19 are the two applicable latest non-CI addresses. Hence, they are part of this calculation. ID#20 is excluded because it is a delete address and not applicable.
$\bar{y}^{(3)}_{non-CI,k}$	2.375	This is the average household size of ID #4 through #6, ID #10 through #11, ID #14 through #15, and ID #18. Note that ID#19 is excluded from the calculation because it is part of the calculation of $\bar{y}_{HHSI,k}$. ID#7 is excluded from this calculation because it is a delete address and not applicable. ID#8,9,13,17 are excluded from this calculation because they are vacant addresses and not applicable.
$\bar{y}_{HHSI,k}$	6	ID #19 is the latest applicable non-CI addresses. Hence, it is part of this calculation. ID#20 is excluded because it is a delete address and not applicable.

Example 2 – Tract-Cell with fewer non-CI addresses than CI addresses

This is an example tract-cell with 3 non-CI addresses and 4 CI addresses (all status imputation addresses).

ID	CI Address	Household (HH) Type	HH Size	Return Date	Status	Occupancy	HH Size
1	Status Imputation					None Exist	None Exist
2	Status Imputation						
3	Status Imputation						
4	Status Imputation						
5	non-CI address	Occupied	7	23-Mar			
6	non-CI address	Occupied	4	24-Mar	Pseudo-CI		
7	non-CI address	Vacant	0	12-Jun	Pseudo-CI		

Using the notation from Attachment A, we give the values for all relevant terms in the tract-cell with an explanation:

Table C.2 – Values for Relevant Terms in Example 2

Term	Value	Explanation
SI_k	2	ID #1 through ID #4 are status imputation addresses. However, we only have 2 non-CI addresses which are reassigned as pseudo-CI addresses (ID #6 and #7).
$OccI_k$	0	None exist
$HHSI_k$	0	None exist
$\bar{y}^{(1)}_{non-CI,k}$	7	Since there are more CI addresses than applicable non-CI addresses, the earliest non-CI address is kept as an average for the non-CI addresses. In this case, ID#5 is kept.
$\bar{y}_{SI,k}$	2	ID #6 and #7 are the two latest non-CI addresses. Hence, they are part of this calculation.
$\bar{y}^{(2)}_{non-CI,k}$	3.667	There are no occupancy imputation addresses so no non-CI addresses are removed.
$\bar{y}_{OccI,k}$	0	There are no occupancy imputation addresses. Hence the value is 0.
$\bar{y}^{(3)}_{non-CI,k}$	5.5	There are no household size imputation addresses so no non-CI addresses are removed for that reason. ID#7 is excluded from this calculation because it is a delete address and not applicable.
$\bar{y}_{HHSI,k}$	0	There are no household size imputation addresses. Hence the value is 0.

Example 3 – Tract-cell with an equal number of non-CI and CI addresses

This is an example tract-cell with 2 non-CI addresses and 2 CI addresses (all occupancy imputation addresses).

					Selected As Pseudo-CI Address by Type		
ID	CI Address	Household (HH) Type	HH Size	Return Date	Status	Occupancy	HH Size
1	Occupancy Imputation				None exist		None exist
2	Occupancy Imputation						
3	non-CI address	Occupied	5	24-Mar			
4	non-CI address	Vacant	0	12-Jun		Pseudo-CI	

Using the notation from Attachment A, we give the values for all relevant terms in the tract-cell with an explanation:

Table C.3 – Values for Relevant Terms in Example 3

Term	Value	Explanation
SI_k	0	None exist
$OccI_k$	1	ID #1 and ID #2 are occupancy imputation addresses. However, we only have 1 non-CI address which is reassigned as a pseudo-CI address (ID #4).
$HHSI_k$	0	None exist
$\bar{y}^{(1)}_{non-CI,k}$	2.5	There are no status imputation addresses so no non-CI addresses are removed for that reason. Hence, the average is determined from ID#3 and ID#4.
$\bar{y}_{SI,k}$	0	There are no status imputation addresses. Hence the value is 0.
$\bar{y}^{(2)}_{non-CI,k}$	5	Since the number of CI addresses and applicable non-CI addresses are equal, the earliest non-CI address is kept as an average for the non-CI addresses. In this case, ID#3 is kept.
$\bar{y}_{OccI,k}$	0	ID #3 is the latest non-CI addresses. Hence, it is the only pseudo-CI address.
$\bar{y}^{(3)}_{non-CI,k}$	5	There are no household size imputation addresses so no non-CI addresses are removed for that reason. ID#4 is excluded from this calculation because it is a vacant address and not applicable.
$\bar{y}_{HHSI,k}$	0	There are no household size imputation addresses. Hence the value is 0.

Example 4 – Tract-cell with one applicable non-CI address and CI addresses that is excluded from the analysis

This is an example tract-cell with 2 non-CI addresses and 1 CI address (household size imputation address).

					Selected As Pseudo-CI Address by Type		
ID	CI Address	Household (HH) Type	HH Size	Return Date	Status	Occupancy	HH Size
1	non-CI address	Occupied	3	24-Mar	None exist	None exist	Pseudo-CI
2	Household Size Imputation						
3	non-CI address	Vacant	0	24-Jun			

In this situation, this cell has a household size imputation. As a result, the only applicable non-CI address is ID#1. ID#3 is not applicable because it is a vacant.

Consequently, $\bar{y}^{(3)}_{non-CI,k}$ and $\bar{y}_{HHSI,k}$ have the same value of 3 and the bias would be 0. As a result, this cell is excluded from the analysis. That means it is excluded as a tract-cell for the computations in Attachment D.

Example 5 – Tract-cell with one applicable non-CI address and CI addresses that is excluded from the analysis

This is an example tract-cell with 4 non-CI addresses and 2 CI addresses (both occupancy imputation addresses).

					Selected As Pseudo-CI Address by Type		
ID	CI Address	Household (HH) Type	HH Size	Return Date	Status	Occupancy	HH Size
1	non-CI address	Delete	0	12-Jun	None exist		None exist
2	Occupancy Imputation						
3	Occupancy Imputation						
4	non-CI address	Occupied	5	24-Jun		Pseudo-CI	
5	non-CI address	Delete	0	25-Jun			
6	non-CI address	Delete	0	26-Jun			

In this situation, this cell has two occupancy imputations. As a result, the only applicable non-CI address is ID#4. ID#5, ID#6, and ID#1 are not applicable because they are deletes.

Consequently, $\bar{y}^{(2)}_{non-CI,k}$ and $\bar{y}_{OccI,k}$ have the same value of 5 and the bias would be 0. As a result, this cell is excluded from the analysis. That means it is excluded as a tract-cell for the computations in Attachment D.

Attachment D

The following formulas explain the statistics in Table 2. The $Bias_k, Variance_k, Bias_M, Variance_M$ terms are defined in Attachment A.

<u>Column (D)</u>

The Tract-cell average squared bias is determined as follows:

$$\text{Tract-cell Average Squared Bias} = \frac{\sum_{k \in \text{Tract-cells}} Bias_k^2}{\#\text{of Tract-cells}}$$

The Merged-cell average squared bias is determined as follows:

$$\text{Merged-cell Average Squared Bias} = \frac{\sum_{M \in \text{Tract-cells}} Bias_M^2}{\#\text{of Tract-cells}}$$

<u>Column (E)</u>

The Tract-cell average variance is determined as follows:

$$\text{Tract-cell Average Variance} = \frac{\sum_{k \in \text{Tract-cells}} Variance_k}{\#\text{of Tract-cells}}$$

The Merged-cell average variance is determined as follows:

$$\text{Merged-cell Average Variance} = \frac{\sum_{M \in \text{Tract-cells}} Variance_M}{\#\text{of Tract-cells}}$$

<u>Column (F)</u>

The Tract-cell average MSE is determined as follows:

$$\text{Tract-cell Average MSE} = \frac{\sum_{k \in \text{Tract-cells}} Bias_k^2 + Variance_k}{\#\text{of Tract-cells}}$$

The Merged-cell average MSE is determined as follows:

$$\text{Merged-cell Average MSE} = \frac{\sum_{M \in \text{Tract-cells}} Bias_M^2 + Variance_M}{\#\text{of Tract-cells}}$$

<u>Column (G)</u>

The % of cells where MSE of the Tract-cell is less than the MSE of the Merged-cell is computed as follows:

For each Tract-cell *k*, the following comparison is made:

- If $Bias_k^2 + Variance_k < Bias_M^2 + Variance_M$ then $\alpha_k = 1$, else $\alpha_k = 0$

Then, we take the average over all the *k:*

$$\text{\% of cells where MSE of the Tract-cell is less than the MSE of the Merged-cell} = \frac{\sum_{k \in \text{Tract-cells}} \alpha_k}{\#\text{of Tract-cells}}$$