Counting Persons Once and Only Once at the Right Location in the Census: Techniques and Challenges Unduplicating People Experiencing Homelessness [1]

Diane F. Barrett and Thomas P. Mc Coy
U.S. Census Bureau, Washington, DC 20233-8700

Abstract

The 2010 Census Service Based Enumeration operation provided people experiencing homelessness an opportunity to be included in the census by conducting the enumeration over a three-day period at shelters, soup kitchens, regularly scheduled mobile food vans and targeted non-sheltered outdoor locations. Since these locations enumerate a transient population, potential exists for a person to be counted at more than one location. Likewise, persons with "no address" who were enumerated at these locations may have also filled out a Be Counted Form. People who were enumerated at soup kitchens and/or mobile food vans could have also been counted at their permanent residence. In an attempt to count people once and only once at the correct location in the census, the Census Bureau conducted an unduplication process using a rule-based and probabilistic matching methodology based on response data to identify duplicate persons. Based on certain predetermined criteria, these duplicates were removed from the final census population count. This paper will discuss the various data processing techniques, challenges and possible research for unduplicating people experiencing homelessness in the census.

Key Words: census, unduplication, people experiencing homelessness

1. Introduction

The 2010 Census Service-Based Enumeration (SBE) was implemented to provide people without conventional housing, including those experiencing homelessness, an opportunity to be included in the census. The SBE was a three-day operation that was conducted March 29, 2010 through March 31, 2010. During the SBE operation, census enumerators visited the following SBE Group Quarters (GQs)[2] emergency and transitional shelters, soup kitchens, regularly scheduled mobile food vans (RSMFVs) and targeted non-sheltered outdoor locations (TNSOLs). Outdoor locations were pre-identified places where people experiencing homelessness lived or stayed without paying to stay which included car, recreational vehicle (RV) and tent encampments.

2. The Service-Based Enumeration Universe

The majority (approximately 94 percent) of the data included in the SBE universe was data collected from Individual Census Reports (ICRs). The ICR was a one-page paper questionnaire used to collect data for an individual during the SBE operation. Enumerators used ICRs to conduct interviews or have the residents fill them out

[1] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

[2] Group Quarters (GQs) is a place where people live or stay, in a group living arrangement that is owned or managed by an entity or organization providing housing and/or services for its residents.

themselves. Enumerators were also allowed to complete these ICRs by other means, that is, via a proxy such as the contact person, use of the facilities' administrative records or by observation.

In addition to the 2010 Census SBE Operation conducted at specific service-based locations, people experiencing homelessness could also participate in the census by completing a Be Counted Form. The Be Counted Form was an unaddressed questionnaire for people who believed they were not counted in the census. The Census Bureau placed these forms at various public locations and other census designated places such as libraries, train stations, Be Counted Sites and Questionnaire Assistance Centers throughout the country during the 2010 Census.

Both the ICRs and the Be Counted Forms collected the same person data, that is first and last name, sex, age, date of birth, Hispanic Origin and race. These forms also asked for other addresses or location information. For the ICR, respondents were asked if they had a usual home elsewhere (UHE), which is a permanent residence other than the service-based location where they were enumerated. If they did, the respondent was to provide the UHE address.

Only a subset of the Be Counted Forms was processed as part of the SBE universe. Those cases where the respondent indicated on the form that they had "no address," that is no permanent residence on April 1, 2010, were asked to provide an address or location information of where they did stay on April 1, 2010. Those cases that met certain conditions were eligible to be included in the SBE universe for further processing.

3. Processing the 2010 Census SBE Universe

Since the 2010 Census SBE Operation was conducted over a three-day period, it was possible, for example, to enumerate a person at a shelter and a soup kitchen and/or mobile food van. It was also possible for a person who was enumerated at a service-based location to fill out a Be Counted Form. As in Census 2000, an attempt was made to unduplicate persons in the SBE universe and count each person once and only once at the right location in the census. The SBE universe was processed in three major steps. The first step was to conduct the unduplication process, then allocate non-duplicate Be Counted Form persons to GQs and finally reconcile persons enumerated at soup kitchens and regularly scheduled mobile food vans against the household composition of their UHE.

3.1 STEP 1 Unduplication Process

3.1.1 Identify the Unduplication Universe

Duplicate records were identified by using standardized demographic data. In order to be included in the unduplication universe all person records from ICRs and the subset of Be Counted Forms (i.e.: those forms that indicated "no address" and the respondent provided location information of where they were on April 1, 2010) had to be coded to at least a state and county.

All person records also had to have sufficient data for matching. The record was required to contain: 1) At least three characters combined in the first and last name fields and 2) At least two person characteristics, one of which was age or date of birth.

3.1.2 Conduct Computer Person Matching to Identify Duplicates,

A rule-based and probabilistic computer matching methodology was used to identify matches, non-matches and possible matches. A final weight was assigned to each pair of person records identified as potential matches. Agreement weights had positive values and disagreement weights had negative values. Variables that were missing from at least one of the two person records involved in the comparison were assigned the weight of zero. A final weight assigned to the pair of person records was the sum of the agreement and disagreement weights for each matching person characteristic (that is, first name, last name, sex, age, date of birth, Hispanic Origin, and race). Matches were determined on whether the first and last name was non-contradictory, the age was within one year of each other and the date of birth was non-contradictory.

Records that were "exact matches" or "exact subset matches" were identified first. A clerical review of possible matches by all stakeholders was conducted to determine the cut off weight. If the cut off weight was greater than or equal to the final weight then all possible matches were considered as duplicates.

Pseudonyms and invalid names: As in Census 2000, enumerators were allowed to enter pseudonyms or invalid names on ICRs when a name could not be obtained. For example, field procedures instructed enumerators to enter the word "Person" in the last name field and a number in the first name field thus the name would appear as "Person 1" on the ICR. However, a clerical review of the last name field in Census 2000 revealed that false duplicates could have resulted when enumerators used pseudonyms to identify SBE residents or when respondents used invalid names in filling out Be Counted Forms. In order to minimize the possibility of erroneously identifying duplicate records, due to the use of pseudonyms and invalid names, certain words such as "Asleep," "Person," "Resident," "Jane or John Doe" were treated as blank during the 2010 Census SBE unduplication process. This process forced the final weight to be based on demographic characteristics rather than the first and/or last name to minimize erroneous duplicates.

3.1.3 Determine the Person Record to Keep as the Survivor

One-person record of the matching pair was selected as the surviving person record by applying specific criteria in the following order:

- *ICR versus Be Counted Form:* If the matching pair was a record from an ICR and the other person record was from a Be Counted Form, then the ICR person record is the survivor and the Be Counted Form person record is the duplicate.
- *UHE versus Non-UHE ICR*: If both records were ICR person records but one record had a UHE and the other record did not indicate a UHE then the ICR with the UHE address is the survivor and the non-UHE ICR person record is the duplicate. The ICR UHE person record was processed further. (See STEP 3, Reconcile Persons at Soup Kitchens and Regularly Scheduled Mobile Food Vans Against the UHE Household Composition.)
- *GQ Type Code:* If both records of the matching pair were Non-UHE ICR person records then the person enumerated with the lowest GQ type code was selected as the survivor. The SBE GQ type codes were as follows:
    - 701 Shelters
    - 702 Soup Kitchens
    - 704 Regularly Scheduled Mobile Food Vans
    - 706 Targeted Non-Sheltered Outdoor Locations

- *Number of demographic characteristics filled*. If the matching pair was from the same type of SBE GQ, then the ICR person record with the most characteristics completed was selected as the survivor.

3.2 STEP 2 Allocate Non–duplicate Be Counted Form Persons to GQs

The next step was to proportionately allocate Be Counted Form persons to GQs. This allowed them to be associated with a Census residential address, which was necessary to tabulate these persons in the census. The allocation took place after the unduplication process for those persons who were enumerated both on an ICR and on a Be Counted Form.

Only those non-duplicate Be Counted Form persons that were assigned a state and county code were retained for allocation. An algorithm was applied proportional to the number of persons in each GQ. We first attempted to allocate Be Counted Persons to occupied SBE GQs (Group 1). If none of these existed in the state and county, then persons were allocated to occupied Adult Group Homes or Residential Treatment Centers (Group 2). If none of these existed in the state and county, then persons were allocated to all other types of GQs (Group 3).

3.3 STEP 3 Reconcile Persons at Soup Kitchens and Regularly Scheduled Mobile Food Vans Against the Usual Home Elsewhere (UHE) Household Composition

The third and final step was to reconcile persons enumerated at soup kitchens and mobile food vans against the UHE household composition. There are persons who have a permanent residence that utilize soup kitchens and mobile food vans. Therefore, it was possible for these persons who were enumerated at these locations during the census to also fill out a regular census household questionnaire. This process was conducted after the final editing for the census and the selection of the representative household return among multiple returns for an address record.

3.3.1 Identify SBE persons with an allowable and valid UHE address.

While all persons could provide an UHE address on the ICR, of the SBE GQs, only those from soup kitchens and mobile food vans were reconciled.

3.3.2 If the UHE was a housing unit address,

If the address was a housing unit, the SBE person was matched[3] against its household composition of the “selected representative “return to determine whether the person was included in the household roster.

- If the SBE person was included in the household roster, then the person was invalidated at the soup kitchen or mobile food van.
- If the SBE person was not in the household roster, for most cases the SBE person was moved into that household and was counted at the UHE address.

[3] The same ruled-based and probabilistic matching and clerical review techniques that were used to identify duplicates in STEP 1 were also used to determine person matches to the UHE roster.

3.3.3 Special cases

There were several special situations where no action was taken. For example, if the UHE address was a GQ, transitory location or a housing unit deleted in field operations, the SBE person was counted at the soup kitchen or mobile food van.

4. 2010 Census Results

4.1 SBE Universe

There were 459,772 person records in the SBE universe data captured from both the ICRs and non-duplicate Be Counted Forms before removing duplicates during the 2010 Census SBE. Approximately 47 percent were data captured at shelters, 45 percent at soup kitchens and mobile Food vans and the remaining eight percent were at outdoor locations.

Table 1. Person Records Data Captured by SBE Type

Type of SBE Service Location	Data Captured Persons	
	Count	Percent of Total
Total	**459,772**	**100.0**
Shelters	214,094	46.6
Soup Kitchens and Regularly Scheduled Mobile Food Vans	207,683	45.2
Targeted Non-Sheltered Outdoor Locations	37,995	8.3

[*]Counts and percentages are unweighted.
[+]Percentages may not sum to 100.0 due to rounding.
Source: Decennial Response File - GQ Person File

4.2 Unduplication Results Between 2010 Census and Census 2000

See Table 2 below that compares the number of persons data captured as well as the duplication rate between the 2010 Census and the 2000 Census SBE. The Census Bureau data captured 201,044 more persons at service-based locations in the 2010 Census (459,772) than in the Census 2000 (258,728).

This increase may have been attributed to the Census Bureau working better with local governments, partnerships and advocacy groups to identify places where people experiencing homelessness stay. In addition, a different universe of people was enumerated in 2010 than in 2000. Due to the economic downturn of 2008, it was likely that more people relied on such services offered by shelters, soup kitchens and mobile food vans in 2010 than in 2000. Changing the definition for outdoor locations in 2010 to include car, RV and tent encampments that formed due to working people who lost their homes when the housing market crashed, may have also had an impact on the difference between the censuses.

Although there was a substantial increase in the number of SBE persons data captured in 2010 over the number in 2000, there were about twice as many more duplicates found in 2000 (six percent) than in 2010 (2.8 percent).

One reason for this difference in the unduplication rate between the censuses may be due to how pseudonyms and invalid names were treated during computer matching. In Census 2000, such names were used in the comparison where the name was assigned a positive or negative weight as appropriate. However, a clerical review of the last name field revealed that, of the 16,787 (six percent of 258,728) duplicates found, about 14 percent of these were most likely false duplicates and were erroneously removed from the census (McNally 2002). As mentioned in section 3.1.2 STEP 1 Unduplication Process, certain pseudonyms and invalid names were treated as missing for 2010 to minimize false positive matches. Variables that were missing from at least one of the two person records involved in the comparison were assigned a weight of zero. This seemed to contribute to the lower unduplication rate for the 2010 Census when compared to the 2000 Census.

Table 2. Service Based Enumeration Unduplication Results

Unduplication	**2010 Census Data Captured Persons**		**Census 2000 Data Captured Persons**	
	Count	**Percent of Total**	**Count**	**Percent of Total**
Total Data Captured Person Records in the SBE Universe	**459,772**	**100.0**	**258,728**	**100.0**
Data Captured Person Records Matched & Not Counted in the Census (Duplicates)	12,774	2.8	16,787	6.0
Total Data Captured Person Records Counted in the Census (*Unique, survivor, or insufficient for matching)*	446,998	97.2	241,094 1	94.0

Counts and percentages are unweighted.

$^+$Percentages may not sum to 100.0 due to rounding.

Source: Decennial Response File - GQ Person & SBE Unduplication Results File

4.3 Persons Added to the SBE Universe from Be Counted Forms

For the SBE universe, there were 26,975 persons added to 7,737 GQs from Be Counted Forms. As expected (see STEP 2), approximately 80 percent were first allocated to SBE GQs where 40 percent were allocated to shelters, 31 percent allocated to soup kitchens and mobile food vans and the remaining nine percent to outdoor locations. Approximately 20 percent were allocated to non-SBE GQ types.

Table 3. Persons Added to the SBE Universe from Be Counted Forms

2010 Census BCF Persons Allocated	BCF Persons		GQs	
	*Count	+Percent of Total	*Count	+Percent of Total
Total BCF Persons	26,975	100.0	7,737	100.0
Total BCF Persons Allocated to SBE GQ	21,681	80.4	6,875	88.9
Shelters	10,835	40.2	3,266	42.2
Soup Kitchens and Regularly Scheduled Mobile Food Vans	8,480	31.4	2,143	27.7
Targeted Non-Shelter Outdoor Locations	2,366	8.8	1,466	19.0
Total BCF Persons Allocated to Other GQs	5,294	19.6	862	11.1

*Counts and percentages are unweighted.

+Percentages may not sum to 100.0 due to rounding.

Sources: 2010 Be Counted Form Geographic Allocation Results File

4.4 Valid UHE Addresses Provided by Respondents Enumerated at Soup Kitchens and Mobile Food Vans

Of the 70,594 persons enumerated on ICRs at soup kitchens and mobile food vans who indicated that they had a UHE, 83 percent provided the address of where they stayed or lived at most of the time. Over half of those that indicated they had a UHE (about 52 percent out of 70,594) provided "valid" addresses that moved on to further processing. (See section 3.3.STEP 3 for the reconciliation against the household composition of the UHE address). The analysis did not provide the number of persons who were counted at their UHE in the census.

Table 4. Valid UHE Addresses Provided by Respondents Enumerated at Soup Kitchens and Regularly Scheduled Mobile Food Vans

Valid UHE Addresses	*Count	+Percent of Total
Total ICRs at Soup Kitchens & Regularly Scheduled Mobile Food Vans that indicated they had an UHE	**70,594**	**100.0**
Total ICRs at Soup Kitchens & Regularly Scheduled Mobile Food Vans that provided a UHE address…………………………………	58,594	83.0
ICRs with valid UHE addresses	36,654	51.9
ICRs with non-valid UHE addresses	21,940	31.1
Total ICRs at Soup Kitchens & RSMFVs without UHE addresses	12,000	17.0

*Counts and percentages are unweighted.

+Percentages may not sum to 100.0 due to rounding.

Source: Decennial Response File Add Tables (rps_address_add_2010)

5. Data Quality Challenges

Although there were many data quality challenges we faced in unduplicating persons experiencing homelessness in the census, the most prevalent challenges were the use of pseudonyms in enumeration and probabilistic matching as well as high item nonresponse rates:

Pseudonyms and invalid names: Enumerators used pseudonyms to identify persons at SBE locations when a first and last name could not be obtained.Although these pseudonyms and invalid names were sufficient to include people experiencing homelessness in the census, they had a low utility during person matching to identify duplicates. In Census 2000, after the unduplication process was conducted, a clerical review of the last name field found that these pseudonyms caused erroneous matches during computer matching... We learned from Census 2000, in order to minimize the possibility of erroneously identifying duplicate records for 2010, certain words such as "Asleep," "Person," "Resident," "Jane or John Doe" were treated as blank for improving the SBE unduplication process during the 2010 Census.

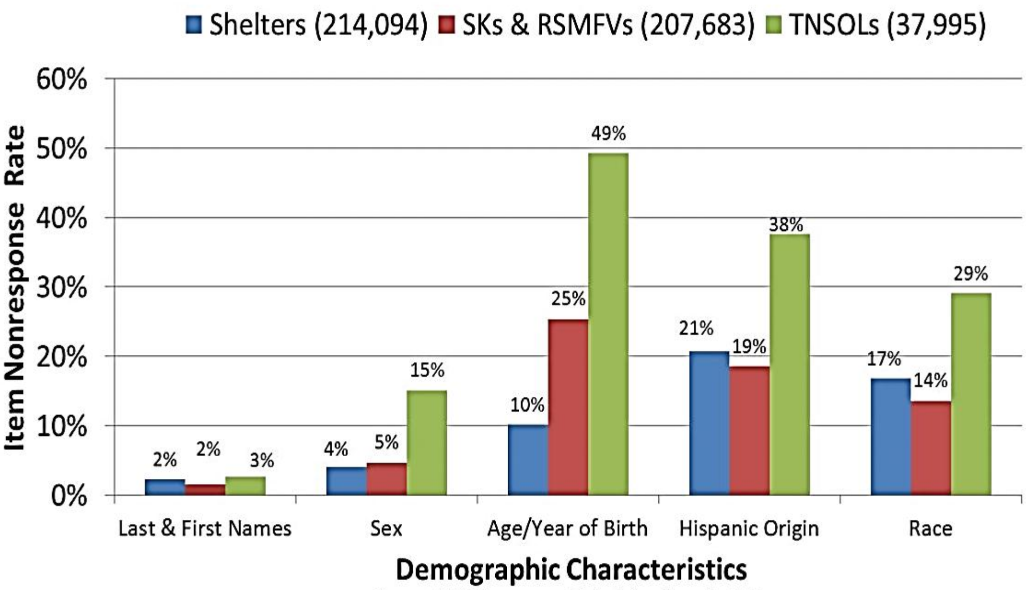
High item nonresponse rates: Another data quality challenge in conducting person matching for this segment of the population was the high item nonresponse rates especially for age, race and Hispanic Origin. Refer to Figure 1 on the next page.

At all three locations, the first and last name fields had a low item nonresponse rate in the 2010 Census. Although there was data entered in these fields, these fields more than likely often contained pseudonyms and invalid names due to enumerator procedures. The analysis did not tabulate how often the last and /or first names fields contained pseudonyms.

Other than requiring a valid name, the age or the year of birth also had to be present to consider that a person record had sufficient data for matching. Note that the item nonresponse rate for age was the most serious for outdoor locations at 49 percent when compared to soup kitchens and mobile food vans at 25 percent and shelters at ten percent.

Although sex, Hispanic Origin and race were not required for a record to have sufficient data for matching, these items in addition to valid name and age are very beneficial in determining strong matches when present. Among the SBE GQs, outdoor locations had the highest item non-response rates for sex, Hispanic Origin and race. Enumeration at outdoor locations was conducted late at night and the early morning hours when most residents were asleep and /or covered up made collecting demographic data by interview or observation very difficult. Because of these high item nonresponse rates, it was almost impossible to unduplicate people at outdoor locations who also may have obtained services at shelters, soup kitchens or mobile food vans during the SBE.

Figure 1. Item Nonresponse Rates by Type of SBE Location[4]



Source: 2010 Census unpublished data from the DRF

6. Moving Forward to the 2020 Census

As part of the strategic goals and objectives for the 2020 Census, the Census Bureau has instituted the 2020 Matching Process Improvement Team to research alternatives and enhancements to improve our current address and person matching methodologies that can be applied to the 2020 Decennial Census operations. Some areas of current research are: evaluating using geographical distance as a matching parameter, collecting and processing more respondent-provided addresses (such as UHEs) and developing Bayesian matching models to assist in automating cut offs for use in probabilistic matching.

Finding ways to reduce the item nonresponse for persons enumerated at SBE GQs, especially those caused by procedural use of pseudonyms in the name field would have a significant impact on data quality. It was also suggested that we standardize and consolidate the SBE unduplication efforts with other census matching processes to improve matching techniques for this segment of the population. For example, incorporate the pseudonyms unique to SBE such as “Asleep,” “Person,” “Diner” and so on into the standard invalid name list used during census matching processes.

7. References

McNally, Tracey (2002), “Census 2000 Evaluation E.6, Revision 1, Service-Based Enumeration,” U.S. Census Bureau, November 6, 2002.

Russell, Deborah and Diane Barrett (2013), “2010 Census Service-Based Enumeration Assessment Report,” U.S. Census Bureau, January 3, 2013.

[4] The numbers in parentheses are the number of data captured persons records at the SBE location. These numbers were used as the denominator to calculate the item non-response rate.