THE EFFECT OF HU/GQ DUPLICATION ON THE AMERICAN COMMUNITY SURVEY[1]

Rob Rothhaas, U.S. Census Bureau, Demographic Statistical Methods Division

KEY WORDS: Housing Unit, Group Quarters, Duplication, Master Address File

Summary

An important source of nonsampling error in frame development is address duplication. This paper examines situations where addresses were included as both housing units (HUs) and group quarters (GQs) in the American Community Survey (ACS) sample frame. Specifically, it discusses the frequency of duplication, the possible sources of duplication, and the impact of duplication on ACS estimates in three sample counties.

Addresses are classified on the Census Bureau's Master Address File (MAF) as both HUs and GQs in situations where they should be one or the other but not both. Minimizing duplication would enhance the ACS sampling universe. The Census Bureau will hopefully take steps to minimize duplication and improve address sources.

Background/Concepts: HU, GQ, ACS, and the MAF

The U.S. Census Bureau divides living quarters into two categories: an HU or a GQ.

An HU may be a house, an apartment, a mobile home, a group of rooms, or a single room that is occupied (or, if vacant, is intended for occupancy) as separate living quarters. Separate living quarters are those in which the occupants live separately from any other individuals in the building and which have direct access from outside the building or through a common hall. People living in HUs make up about 97% of the U.S. population.

The Census Bureau classifies persons not in HUs to live in GQs and recognizes two general categories of people in GQs: institutionalized population and noninstitutionalized population. The institutionalized GQ population includes people under formally authorized, supervised care or custody in institutions. The noninstitutionalized GQ population includes people who live in GQs other than institutions like college dormitories and military barracks.

GQ type codes and descriptions appear in Census 2000 Summary File 1 technical documentation at http://www.census.gov/prod/cen2000/doc/sf1.pdf.

GQs are located in "special places." Special places are locations that the Census Bureau identifies as possible sources of GQs for the Decennial GQ universe. An example of a special place is a university. Dormitories, fraternity houses, and sorority houses on and off campus are GQs that the Census Bureau associates with the university, the special place.

ACS is mandated by Title 13, Sections 182 and 225 of U.S. Code. It is intended to replace the Decennial Census long form. In addition to age, race, sex, and ethnicity data that the short form gathers, the long form generates detailed data about concepts like employment, journey to work, and income.

Most of the U.S. population received the short form in Census 2000. About one of every six housing units received the long form. In GQs, Census 2000 enumerators attempted to distribute the long form to about one of every six persons. For GQs, the goal was for residents to either fill out the long form or provide responses via enumerator interview.

ACS enumerates a sample of persons every month. Instead of having to wait ten years to release data, ACS regularly generates estimates of long form characteristics. These estimates are available annually for big cities and other large geographic areas. Multi-year estimates will supply data for smaller geographic entities, for which samples spread over several years are necessary to provide sufficient reliability.

ACS HU enumeration has taken place since 1996. Each year, the Census Bureau releases official ACS HU data for many parts of the country. ACS GQ enumeration has taken place since 1997. However, the Census Bureau has not yet released detailed ACS GQ data. Due to insufficient funding, ACS GQ enumeration is not taking place in 2002.

The source of ACS sample is the MAF: an inventory of U.S. addresses that the Census Bureau's Geography Division has created and maintains. The MAF is also the source for the addresses that the Census Bureau used to conduct Census 2000.

Addresses enter the MAF via many different sources. The base starting point for the MAF was the 1990 Census Address Control File (all addresses collected in the 1990 Census). Address listing operations that the Census Bureau conducted for Census 2000 identified new addresses for the MAF as well as updates to existing addresses. Since Census 2000, the Census Bureau regularly receives updated Delivery Sequence Files (DSFs) from the U.S. Postal Service that identify mail delivery points. Geography Division

[1] This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

applies DSF updates to continuously update the MAF. DSF updates allow ACS to include in its sampling universe new construction addresses that did not exist at the time of prior Decennial Censuses. The current source for MAF GQ addresses is Census 2000 GQ operations. Other Census Bureau operations will also feed into MAF HU and GQ updating.

Why HU/GQ Duplication is Problematic

Address duplication in the sample frame, or universe, is problematic for demographic surveys like ACS. It causes overcoverage because the same address inappropriately appears multiple times. Misidentification of HUs as GQs and GQs as HUs causes an address to have at least two chances of selection in situations where it should only have one chance, or even no chance. Persons in GQs tend to have different demographic characteristics than persons in HUs, like lower incomes. Therefore, misidentification of HUs as GQs and GQs as HUs could negatively impact the aggregate data.

Since the GQ population is much smaller than the HU population, survey designers wishing to make separate GQ estimates (as may eventually be the case for ACS) may find address duplication more troubling for GQ estimates than for HU estimates.

Example of Possible HU/GQ Duplication

Here is an example of possible address duplication on the MAF. A basic address (where a basic address is a house number and street name ignoring the unit designation) has 252 Census 2000 HU records with a total Census 2000 population of 298. Each of these records represents a different apartment at the address. The same basic address appears as GQ college housing on the MAF with a Census 2000 population of 297. In reality, the address is the location of an apartment tower where a college houses students. The college's web site says of the tower, "The 252 apartments accommodate approximately 300 students and their families." These counts suggest the possibility of address duplication between the HU and GQ records.

For some address duplication scenarios, the address is really a GQ. Situations also occur where HUs tend to look like GQs. An example is independent living units in an assisted living facility where residents do not receive twenty-four hour nursing care. The Census Bureau classifies these units as HUs.

Reasons for HU/GQ Duplication

Address duplication can occur for many reasons. An important reason is the changing nature of the GQ universe. Formerly, GQs often existed in large special places that were set off from HUs in the surrounding neighborhood. Special places were often campus settings that made the Census Bureau's distinction between HUs and GQs readily apparent. GQs less frequently fit this paradigm today.

Many persons living in GQs have physical impairments, infirmities, or addiction problems. Past societal trends tended to institutionalize these persons. Current trends assimilate them into the population more readily. GQ housing where these kinds of people often live more and more tends to resemble HUs.

Government regulations are increasingly dictating that persons in these situations must live in small groups of some maximum size instead of larger, more institutional settings. For example, a group home may just have a few persons living in it and may look from the outside like an everyday single family house. The Census Bureau considers a group home to be a GQ. Because many resemble HUs so closely, HU records for group home addresses tend to appear on the MAF.

Also making this problematic is the fact that these kinds of GQs now tend to be throughout residential communities. No longer do GQs so frequently cluster together in a separate special place.

HUs and GQs can legitimately share an address at a special place. One such scenario is when HUs are "embedded" within GQs. An example of an embedded HU is an apartment for a dormitory manager within a college dormitory. The dormitory is the GQ and the dormitory manager's apartment is an HU embedded within a GQ. Another example of a situation that could cause an HU to have the same address as a GQ is an HU "freestanding" in a special place, such as a university president's house on a college campus.

While HUs and GQs can share an address, the danger exists of HU and GQ addresses on the MAF duplicating each other when the Census Bureau misinterprets duplication of the address by thinking it is due to the presence of embedded or freestanding HUs. That is, an address could enter the MAF through an HU operation even though the address also entered the MAF through a Census 2000 GQ operation. Suppose an address is either an HU or GQ but not both. If HU and GQ records exist on the MAF for the same address, this is address duplication that is inappropriate.

ACS Sampling

The ACS HU sampling universe consists of "potentially good" HU records on the MAF. "Potentially good" means that the record purportedly represents a residential HU that Geography Division has not flagged as a delete on ACS MAF extracts.

ACS HU sampling occurs in two stages. For the first stage, each HU universe record within a county has an equal chance of selection except for the provision that an HU can only be in the ACS sample once every five years. Second stage HU sampling is dependent upon the size of functioning governmental units. Small functioning governmental units have higher second stage sampling rates than large ones to provide sufficient sample for estimation.

ACS GQ sampling takes place once for a calendar year. The sampling identifies the sample for the next calendar year. For the most part, the current source of the ACS GQ universe is GQs that the Census Bureau enumerated for Census 2000.

For ACS, the Census Bureau divides the GQ universe into “small” and “large” GQs. Small GQs have a GQ population of fifteen or less. Each small GQ has an equal chance of selection except for the provision that a small GQ can only be in the ACS sample once every five years. ACS typically enumerates all residents of small sample GQs. Large GQs have a GQ population greater than fifteen and are sampled with probability proportionate to size. Each large GQ hit receives a within-GQ sampling pattern whose goal is to enumerate an expected ten residents.

MAF Extracts That This Analysis Uses

This analysis uses ACS MAF extracts (specified subsets of the complete MAF) that Geography Division provided for the main phase of ACS HU sampling for 2002. They include the HUs and GQs needed for analysis.

Data Analysis: General Methodology

This analysis looks first at the overall ACS sampling frame, then specifically at the 2002 sample within that frame.

It focuses on three counties: Broward County, Florida; Bronx Borough, New York; and Harris County, Texas. They are in both the 2002 ACS HU sampling universe and the 2002 ACS GQ sampling universe. The 2002 ACS GQ sampling universe consists of thirty-six counties chosen to compare Census 2000 data to ACS data.

Broward County, The Bronx, and Harris County have large populations that are demographically diverse. They have sizable GQ populations. Because the GQ universe is very small, especially compared to the HU universe, use of large counties for this analysis is essential to obtain sufficient data.

A “city style” address has a standard house number and street name. An example is “101 Main Street.” Non-city style addresses include those that consist of rural route and box numbers, post office boxes, or location descriptions. City style addresses are preferable for identifying possible HU/GQ duplication because non-city style addresses often do not identify physical locations and, especially for location descriptions, can be vague and wordy. Most addresses in Broward County, The Bronx, and Harris County that identify physical locations are city style.

This analysis uses a computer method to detect possible HU/GQ duplication. This method analyzes strings of characters comprising addresses. It considers those strings that are identical within a county for both HU and GQ MAF records to be possible duplicates.

The main advantages to a computer method versus a manual method are lower cost, ability to replicate, and time savings. Since addresses on the MAF are standardized, computer use was efficient.

The computer method has disadvantages. If addresses have minor differences on the MAF that do not actually constitute separate addresses, the computer method will consider them different addresses and fail to identify them as possible duplicates. This is true for MAF addresses that include typographical errors and that use alternate wording or spellings.

This analysis only identifies possible HU/GQ duplication due to HUs and GQs having the same address. This analysis includes no followup to determine if possible duplication is actual duplication.

This analysis only considers MAF HU records that the Census Bureau considers to be “potentially good” and thus eligible for the ACS HU sampling universe. It only considers MAF GQ records that had positive integer final GQ populations in Census 2000.

A flag exists on the MAF to identify HUs embedded in GQs. This analysis excludes as possible duplicates addresses that otherwise qualify but whose HUs all appear as embedded HUs on the MAF.

This analysis may sometimes identify HUs and GQs legitimately sharing an address as possible duplicates even though they are not.

Data Analysis: Computer Criteria

The computer method that this paper uses analyzes address duplication two ways. One is through “tight” criteria that forces HU and GQ addresses to match exactly to qualify as possible duplicates. The other is through a broader, “loose” criteria that is less rigorous in how it requires HU and GQ addresses to match to qualify as possible duplicates.

Each criteria attaches several MAF address fields and uses the SAS “compress” function to remove intermediate blanks. It requires the resulting character string to exactly match between one or more HU records in a zip code and one or more GQ records in a zip code to qualify as a possible duplicate.

The MAF fields that the criteria attaches are: house number prefix; house number 1; house number separator; house number 2; house number suffix; street prefix direction; street prefix type; street name; street suffix type; street suffix direction; street extension; and within structure identifier. The first five comprise the house number. The next six comprise the street name. The last is part of the unit designation.

The MAF identifies a unit designation by using a within structure description and identifier. If the unit designation is "APT 1," the within structure description is "APT" and the within structure identifier is "1." Neither criteria uses within structure description because it could exclude too many actual duplicates.

The tight criteria considers HUs and GQs to possibly duplicate if they share a house number, street name, and unit designation in a zip code. The loose criteria considers HUs and GQs to possibly duplicate if they share a house number and street name, regardless of unit designation, in a zip code.

For example, suppose the address for an HU in a zip code is "101 MAIN ST APT 1." Suppose the address for a GQ in the same zip code is "101 MAIN ST." These addresses are not duplicates under the tight criteria. The tight criteria considers the unit designations ("APT 1" and blank) to differ. The addresses are duplicates under the loose criteria since only "101 MAIN ST" is the comparison string.

In the computer match, an address can only qualify as a possible duplicate if its house number and street name are nonblank. House number 1 and street name are the MAF fields used to make this determination. Thus, only city style addresses can be possible duplicates in this analysis. The criteria do not look at rural route, box, or location description fields.

Data Analysis: Caution

Before percentages are presented, the reader is cautioned to take care interpreting them.

Structures can be in denominators of percentages as members of the sample frame but will never be in numerators as duplicates. Examples are incomplete addresses, non-city style addresses, addresses that have incorrect or alternate spellings, and embedded HUs. The computer method never classifies these units as duplicates.

Some structures would never be expected to duplicate. An example is a large prison, which is a special place with many GQs. One would not expect a large prison to be mistakenly classified as HUs. Residential HUs without characteristics making them eligible for the GQ universe would not be expected to have GQ records duplicating HU records on the MAF.

As a result, duplication percentages will never be close to 100%. Any duplication that is real duplication is undesirable. Beyond these cautions, interpretations made by the reader about whether the following duplication percentages are "high" or "low" are subjective.

Data Analysis: Results for 2002 ACS Sample Frame

Table 1 shows possible HU/GQ duplication using the tight and loose criteria as percentages of duplicate GQs to total GQs and GQ duplicate population to total GQ population.

Table 1 - Possible HU/GQ Duplication in Sample Frame

	Tight		Loose	
County	**GQs**	**GQ Pop**	**GQs**	**GQ Pop**
Broward	6%	2%	16%	15%
Bronx	17%	7%	37%	17%
Harris	7%	3%	14%	9%
Total	9%	4%	21%	13%

Less than 1 percent of HUs are potentially duplicated structures using either criteria. The rounded percentage is not less than 1 percent under the loose criteria for The Bronx. Less than 1 percent of all living quarters are potentially duplicated structures using either criteria. The rounded percentage is not less than 1 percent under the loose criteria for The Bronx.

Less than 1 percent of the Census 2000 HU population lives in potentially duplicated structures using either criteria. The rounded percentage is not less than 1 percent under the loose criteria for The Bronx. Rounded, one percent of the total Census 2000 population lives in potentially duplicated structures using the loose criteria. By county, the percentage is 2 percent for The Bronx and less than 1 percent for Broward and Harris Counties.

GQ types with the most potentially duplicated GQs using the tight criteria are noninstitutional group homes/halfway houses (48% of GQ duplicates), religious GQs (19%), and residential care facilities providing "protective oversight" (13%).

GQ types with the most potentially duplicated GQs using the loose criteria are noninstitutional group homes/halfway houses (38%), religious GQs (14%), and homeless shelters (13%).

GQ types with the highest Census 2000 population of residents living in potentially duplicated GQs using the tight criteria are college housing (39% of GQ duplicate population), homeless shelters (24%), and noninstitutional group homes/halfway houses (14%).

GQ types with the highest Census 2000 population of residents living in potentially duplicated GQs using the loose criteria are homeless shelters (24%), college housing (19%), and nursing homes (19%).

Nationwide, college housing, residential care facilities providing "protective oversight," and noninstitutional group homes/halfway houses are among the GQ types with the highest percentages of HUs in the same special places as GQs. Religious GQs, homeless shelters, and nursing homes are among the GQ types with the lowest percentages of HUs in the same special places as GQs. This means that college housing, residential care facilities providing "protective oversight," and noninstitutional group homes/halfway houses exhibit much possible HU/GQ duplication while also exhibiting many instances of legitimately separate HUs and GQs in close proximity. Religious GQs, homeless shelters, and nursing homes exhibit much possible HU/GQ duplication but not many instances of legitimately separate HUs and GQs in close proximity.

According to a preliminary report on the Census 2000 Housing Unit Coverage Study, small multi-unit addresses had higher rates of "erroneous enumerations" in Census 2000 than either single unit addresses or large multi-unit addresses. The amount of possible HU/GQ duplication for noninstitutional group homes/halfway houses seems to be consistent with this finding (Report 17 of Executive Steering Committee for Accuracy and Coverage Evaluation Policy II, "ESCAP II: Census 2000 Housing Unit Coverage Study," October 17, 2001).

Data Analysis: Impact on ACS 2002 Sample

The Census Bureau canceled 2002 ACS GQ. However, the Census Bureau selected sample for 2002 ACS GQ before the cancellation. The sampling rate was 2.5 percent across the thirty-six ACS comparison counties. The author analyzed the sample and applied baseweights to it to derive 2002 ACS GQ weighted sample results.

For 2002 ACS, 10 of 226 sample GQs in Broward County, The Bronx, and Harris County are potentially duplicated structures using the tight criteria. Thirty-nine of the 226 GQs are potentially duplicated structures using the loose criteria. Table 2 breaks down numbers of sample GQs in potentially duplicated structures by whether the GQs were small or large.

Table 2 – 2002 ACS GQ: Numbers of Possible HU/GQ Duplicate Sample GQs (Base=226 GQs)

County	Tight		Loose	
	Small (34)	Large (192)	Small (34)	Large (192)
Broward	1	0	3	6
Bronx	2	3	4	15
Harris	2	2	2	9
Total	**5**	**5**	**9**	**30**

To assess the impact of possible HU/GQ duplication on the ACS 2002 sample population, the author derived weighted 2002 ACS GQ information as follows in the absence of 2002 ACS GQ survey data. A 2.5 percent sampling rate yields a baseweight of 40, where 40 equals 1 divided by .025.

ACS aims to enumerate all residents of small sample GQs. To derive weighted information for small GQs, the author multiplied the GQ population input into sampling by 40 for each small sample GQ and summed across all small sample GQs. This yields a weighted estimate of the small GQ population.

The ACS goal for large sample GQs is to enumerate ten GQ residents per hit, where a large sample GQ has one or more hits depending on its size and a systematic sampling pattern. To derive weighted information for large GQ hits, the author multiplied 10 by 40 to yield a hit weight of 400. Applying this weight to each hit (where a large sample GQ can have one or more hits) and summing across hits yields a weighted estimate of the large GQ population.

For a county, summing the weighted estimates of the small and large GQ populations yields an estimate of the total GQ population.

Table 3 shows the percentage of total ACS GQ weight consisting of residents of GQs in potentially duplicated structures for the tight and loose criteria.

Table 3 –Possible HU/GQ Duplicates in 2002 ACS Weighted GQ Population

County	Tight	Loose
Broward	2%	18%
Bronx	3%	15%
Harris	2%	9%
Total	**3%**	**13%**

Using the tight criteria, all of the duplicate GQ weight in Broward County comes from noninstitutional group homes/halfway houses. For The Bronx, duplicate GQ weight comes from homeless shelters (59%), college housing (29%), and noninstitutional group homes/halfway houses (12%). For Harris County, duplicate GQ weight comes from college housing (77%) and noninstitutional group homes/halfway houses (23%).

Using the loose criteria, the most duplicate GQ weight for Broward County comes from residential care facilities providing "protective oversight" (35%), nursing homes (24%), and homeless shelters (18%). For The Bronx, the most duplicate GQ weight comes from homeless shelters (35%), noninstitutional group homes/halfway houses (30%), nursing homes (12%), and college housing (12%). For Harris County, the most duplicate GQ weight comes from nursing homes (31%), college housing (31%), and residential care facilities providing "protective oversight" (21%).

Eliminating HU/GQ Duplication

The Census Bureau has many possible ways to minimize HU/GQ duplication. There are methods currently in place to address this issue and the author suggests hereinafter others that could be used in the future.

The Count Question Resolution (CQR) program is an administrative review program that handles external challenges to particular official Census 2000 counts of HU and GQ population received from state, local, or tribal officials of governmental entities or their designated representatives. CQR identifies some instances of HU/GQ duplication and removes the duplication from the MAF. However, CQR's limited scope does not allow wholesale resolution of HU/GQ duplication nationwide.

The Census Bureau's Local Update of Census Addresses (LUCA) program also affords opportunities for the Census Bureau to work with local governments to potentially eliminate MAF duplicates.

The Census Bureau has been eliminating some HUs from its ACS HU sampling universe even though the MAF identifies them as "potentially good" HUs. This has mostly resulted from an ACS GQ operation that has taken place at the Census Bureau's National Processing Center (NPC) in Jeffersonville, Indiana to identify possible imperfections in Census 2000 living quarters information. This operation has only taken place in the thirty-six ACS GQ counties and has not been a comprehensive review of the ACS sampling frames in these counties. The operation is not currently taking place due to the cancellation of 2002 ACS GQ enumeration. The operation consists of NPC identifying some MAF HU records that have the same addresses as GQs in sample ACS special places. When NPC encounters this situation, NPC notifies Census Bureau Headquarters. Headquarters then performs MAF and Internet research to attempt to determine whether HU/GQ duplication exists and, if so, whether addresses should be HUs or GQs. If Headquarters cannot tell, Headquarters sometimes authorizes NPC to contact the special place to find out more details about the living arrangements there. If Headquarters concludes that address duplication exists and that an address should be one or more HUs, not one or more GQs, Headquarters flags the GQ records as ACS GQ universe deletes on an ACS database. If Headquarters concludes that address duplication exists and that the address should be one or more GQs, not one or more HUs, a Headquarters operation takes place to remove the MAF HU records from the ACS HU universe. Currently, no system exists to flag duplicates that the NPC operation identifies as deletes on the MAF itself.

General approaches for the future include reconsidering Decennial HU and GQ definitions to more clearly distinguish between what living quarters should qualify as HUs versus GQs. The Census Bureau should also consider a higher degree of integration between Decennial HU and GQ operations and more rigorous unduplication operations.

The MAF criteria for adding DSF updates could be more stringent to ensure that DSF HU adds are actually HUs instead of GQs. This would not be easy or foolproof.

Census Bureau field listing operations can identify instances of duplication and resolve them. Through these operations, field listers could determine whether addresses exhibiting HU/GQ duplication are actually HUs versus GQs and then delete improperly duplicated records.

The Census Bureau should perhaps consider removing from its ACS HU sampling frame MAF HU records that exhibit possible HU/GQ duplication and are not flagged as embedded within GQs. A drawback to this approach is that it would result in the removal of some addresses that exhibit HU/GQ duplication and should really be HUs instead of GQs.

Other ways exist to identify and resolve MAF HU/GQ duplication. Examples include using administrative records and business establishment lists as independent sources to compare to the ACS frames.

Conclusion

The MAF classifies some addresses as both HUs and GQs. Frequently, the addresses should consist of either HUs or GQs but not both. Minimizing this duplication would enhance the ACS universe and the ultimate quality of the resulting data.