RESEARCH REPORT SERIES
(Statistics #2010-09)

Imputation Procedures for American Community Survey Group Quarters Small Area Estimation

Chandra Erdman
Chaitra H. Nagaraja

Statistical Research Division
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: December 21, 2010

Disclaimer: This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

Imputation Procedures for American Community Survey Group Quarters Small Area Estimation

Chandra Erdman and Chaitra H. Nagaraja
U.S. Census Bureau

December 21, 2010

Abstract

The American Community Survey (ACS) sampling procedure for group quarters facilities is designed to obtain state-level estimates. However, estimates for the total resident population – household and group quarters residents combined – are published at the census tract and county levels as well. A consequence of the sampling procedure is that often little to no group quarters data are collected for small areas. To improve estimates at lower levels of geography, four imputation procedures are proposed. The objective is to obtain, through sampling or imputation, county by major type group representation for 1- and 3-year estimates and tract by major type group representation for 5-year estimates. Entire person records are imputed, using the sampled group quarters records, into areas with little sample coverage. Donors are chosen based either on geographic proximity or the similarity of the surrounding housing population. To evaluate each imputation procedure, a group quarters population is simulated using Census 2000 short-form data. From this population, twenty-five independent ACS samples are drawn and each imputation procedure is applied to the simulated samples. The resulting 1-, 3-, and 5-year imputation-based estimates of basic demographic variables are compared to the design-based estimates and to the simulated population values. The bias of the imputation-based estimates is found to be larger than the design-based estimates, but the variance is smaller.

Keywords: American Community Survey, group quarters, small area estimation, imputation, cluster analysis.

Disclaimer: This report is released to inform interested parties of ongoing research and to encourage discussion. The views expressed on statistical issues are those of the authors and not necessarily those of the U.S. Census Bureau.

1 Introduction

The American Community Survey (ACS) is a replacement sample survey for the Census long-form and covers all 50 U.S. states and the District of Columbia. Data are collected continually throughout the year on a diverse set of characteristics including education, income, housing, and demographics. Annual estimates are published for geographies in which more than 65,000 people reside, 3-year estimates are published for areas in which more than 20,000 people live, and 5-year estimates are produced for all Census block groups. Sampling is performed separately for households and group quarters (GQ) facilities because residents of these two sets of people tend to have very different characteristics. At the state and national levels, the U.S. Census Bureau releases select characteristics of the GQ population. However, at sub-state levels, such as county and census tract, data from GQ residents is combined with household residents and estimates are published for the total resident population.

The designation "group quarters" encompasses a wide range of facilities from prisons to college dormitories. All GQ facilities are classified into one of seven major GQ types[1]: (1) Correctional Institutions, (2) Juvenile Detention Facilities, (3) Nursing Homes, (4) Other Long-term Care Facilities, (5) College Dormitories, (6) Military Facilities, and (7) Other Non-institutional Facilities. Each facility is further classified into one of nearly 30 specific GQ types (listed in Table 1). Major types 1-4 are considered institutional GQs and types 5-7 are non-institutional. Every year since 2006, the American Community Survey has sampled approximately 2.5% of the expected number of persons residing in group quarters facilities. GQs are divided into one of two strata for sampling purposes based on the expected size of the GQ: small (15 or fewer residents) or large (more than 15 residents). According to the U.S. Census Bureau (2009), there are approximately 105,000 small GQs, 77,000 large GQs, and only 3,000 GQs which have an unknown size. Small group quarters facilities are eligible for sampling only once in a 5-year period and, if selected in sample, all residents are interviewed[2]. Large group quarters facilities are sampled according to a two-stage clustered design. When the ACS was first implemented for group quarters, residents of qualifying facilities in each state were sampled at a rate of approximately 1/40; researchers found, however, that this rate was often too low to obtain quality estimates and was increased for 15 states starting in 2008. Further details can be found in Section 2.2.

The state-level sampling design leads to many sub-state areas which have few or no GQs in sample. Consequently, estimates for many small and/or less populous areas do not accurately reflect the group quarters population residing there. Estimates published for the total population (household residents combined with GQ residents) often vary significantly from year to year simply because GQs are in sample in one year and not in the next. Beaghen and Stern (2009) provide specific examples of such variability.

In order to produce non-zero (and hopefully more reasonable) estimates of the group quarters population at smaller levels of geography, we plan to impute whole person records into selected facilities which are not in sample. We propose four whole-record imputation methods comprised of two ways of selecting which GQs to impute into combined with two ways of selecting donors for imputation. Our primary goal is to provide non-zero 1-, 3- and 5-year estimates of the GQ population for each combination of county and major GQ type group on the sampling frame. Our secondary goal is to provide 5-year estimates for all combinations of tract and major GQ type group on the sampling frame. To examine the effect of this large-scale imputation, each imputation method is applied to simulated ACS samples based on Census 2000 short-form data. This work is the initial phase of a larger group project on GQ small area estimation for the ACS. If the results from the testing phases are favorable, the imputation methodology outlined in this paper will be incorporated into the 2011 ACS 1-, 3-, and 5-year estimation production process.

This paper is organized as follows. In Section 2 the simulation study is described and in Section 3 each imputation method is outlined. We conclude with results in Section 4 and discussion in Section 5.

Table 1: ACS GQ Type Codes and Definitions

(1) Correctional Institutions	
(101)	Federal Detention Centers
(102)	Federal Prisons
(103)	State Prisons
(104)	Local Jails and Other Municipal Confinement Facilities
(105)	Correctional Residential Facilities
(2) Juvenile Detention Facilities	
(201)	Group Homes for Juveniles

[1] We use the terms "major GQ type" and "GQ type" interchangeably here

[2] As long as there are 15 or fewer residents at the time of interviewing.

(202)	Residential Treatment Centers for Juveniles
(203)	Correctional Facilities Intended for Juveniles
(3) Nursing Homes	
(301)	Nursing Facilities/Skilled Nursing Facilities
(4) Other Long-term Care Facilities	
(401)	Mental (Psychiatric) Hospitals/Psychiatric Units in Other Hospitals
(402)	Hospitals with Patients Who Have No Usual Home Elsewhere
(403)	In-Patient Hospice Facilities
(404)	Military Treatment Facilities with Assigned Patients
(405)	Residential Schools for People with Disabilities
(5) College Dormitories	
(501)	College/University Housing
(6) Military Facilities	
(106)	Military Disciplinary Barracks and Jails
(601)	Military Quarters
(602)	Military Ships
(7) Other Non-institutional Facilities	
(701)	Emergency and Transitional Shelters for People Experiencing Homelessness
(702)	Soup Kitchens, Mobile Food Vans, Non-Sheltered Outdoor Locations*
(703)	Domestic Violence Shelters*
(801)	Group Homes Intended for Adults
(802)	Residential Treatment Centers for Adults
(900)	Crews of Maritime Vessels*
(901)	Workers Group Living Quarters and Job Corps Centers
(902)	Religious Group Quarters
(903)	Natural Disaster*

**Specific type not sampled by the ACS.*

2 Simulation Study

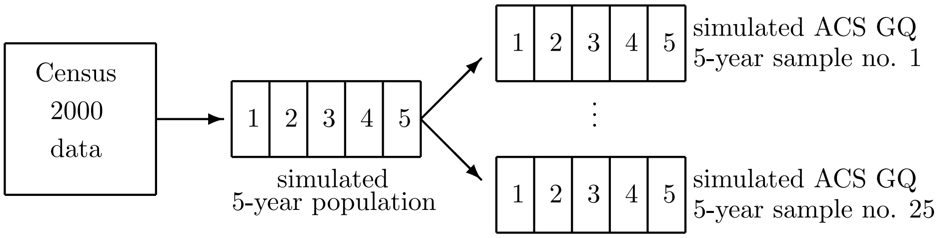
For this study, ACS GQ samples are drawn from a population simulated using the Census 2000 short-form data for each state and the District of Columbia. To generate these sampled records, we follow the ACS sampling procedure outlined by the U.S. Census Bureau (2009) and the ACS group quarter sampling specifications. We attempt to make the process as realistic as possible; however, we do make some simplifications. To this end, we allow the population size of each GQ to vary across years and we simulate a limited form of GQ closings. In total, we simulate one 5-year population set using the Census 2000 data and draw 25 sets of 5-year samples from that population. Figure 1 is a diagram of this procedure.

To evaluate estimates made from both the current and proposed methodologies, we compare them to the simulated population values, which we consider as the “true” values. To evaluate 1-year estimates, we use the fifth year of each five year sample, for 3-year estimates, years three through five are used, and, for 5-year estimates, all five years of each sample are used.

2.1 Simulating the GQ Population

To imitate what happens in reality, we simulate a new GQ population each year (1-5). An algorithm, developed by Joyce (2010) is used to set the size of each GQ in the universe for each year. The number of residents in each GQ can change *across* years, but stays constant *within* a year. Then,

Figure 1: Simulation Procedure



GQs are deleted from the GQ universe based on rates computed using actual ACS data. Finally, each GQ is populated by drawing persons from the corresponding Census 2000 records for that GQ. The residents are fixed within a year but a new set of residents are drawn each year. To describe this procedure, we begin with some notation which is also summarized in Table 2. For each GQ i of major type k, in county c, tract t and in year y:

1. $y = 0, 1, 2, 3, 4,$ or 5 where $y = 0$ indicates the Census 2000 time period and $y = 1, 2, \ldots$ indicate the years when samples are taken.

2. $o_{i,k,c,t,y}$ is the current population. For instance, $o_{i,k,c,0,y}$ is the population of GQ i in Census 2000.

3. $e_{i,k,c,t,y}$ is the expected population of GQ i in year y. This construct is only relevant for years when samples are taken, not for Census 2000. When constructing the true population, $e_{i,k,c,t,y} = o_{i,k,c,t,y-1}$. When selecting samples, $e_{i,k,c,t,y}$ is the most recent observed current population value (see Section 2.2 for more specific details).

4. GQs are assigned each year to one of two strata: small (15 or fewer residents) or large (more than 15 residents) based on the expected population, $e_{i,k,c,t,y}$.

Table 2: General Notation

Symbol	Definition
s	state
c	county
t	tract
y	year (Census 2000 is year 0)
k	GQ major type
i	GQ
j	person
r_s	sampling fraction for state s
$e_{i,k,c,t,y}$	expected number of persons in GQ i of major type k which is in county c, tract t in year y
$o_{i,k,c,t,y}$	observed number of persons in GQ i of major type k which is in county c, tract t in year y

For each year (1-5), the population from which we sample is simulated according to the following algorithm.

Population Construction Algorithm

1. If $y = 1$, compute the current population, $o_{i,k,c,t,1}$, for each GQ as a function of the expected population $e_{i,k,c,t,1}$. If $y > 1$, only compute the current population, $o_{i,k,c,t,y}$, for each large GQ using $e_{i,k,c,t,y}$, the expected population[3]. Recall that $e_{i,k,c,t,y} = o_{i,k,c,t,y-1}$ when simulating the population.
2. Apply the GQ delete rates, computed based on year of sample, major GQ type group (1-7), and the size determined using the expected population count ($e_{i,k,c,t,y}$). The rates for year 1 correspond to data analyzed from 2006, year 2 from 2007, and year 3 from 2008. Data on deletion rates did not exist for 2009 or 2010 at the time of this research so 2008 rates are also applied to years 4 and 5. If the delete rate is x percent, then x percent of the GQs in the universe with the specified major type group and size combination are randomly selected to be deleted. For deleted GQs, set the current population count, $o_{i,k,c,t,y}$, to 0, indicating that the GQ no longer exists in the current and subsequent years. These rates were calculated by P. Joyce (2010).
3. For each GQ i which exists in year y, choose person records from Census 2000 to populate the GQ. Records are chosen as follows:
   (a) If $o_{i,k,c,t,y} \leq o_{i,k,c,t,0}$, choose $o_{i,k,c,t,y}$ persons without replacement from $o_{i,k,c,t,0}$.
   (b) If $o_{i,k,c,t,y} > o_{i,k,c,t,0}$, let $n = \lfloor o_{i,k,c,t,y}/o_{i,k,c,t,0} \rfloor$ where $\lfloor \cdot \rfloor$ is the floor function. Then take n copies of all $o_{i,k,c,t,0}$ persons from the census. Finally, select $o_{i,k,c,t,y} - no_{i,k,c,t,y}$ persons without replacement from the $o_{i,k,c,t,0}$ census persons. Add these persons to the initial list of $no_{i,k,c,t,y}$ people to obtain the full population of that GQ.
4. Set the expected population for the next time period to the current population: $e_{i,k,c,t,y+1} = o_{i,k,c,t,y}$.

2.2 Construction of Samples

The ACS sampling methodology is used to choose samples for both small and large stratum GQs. Since we are attempting to mimic the actual ACS sampling and data collection process in our simulation, GQ types which are not sampled in the ACS are omitted. To reduce the complexity of the simulation, we assume that all sampled persons are survey respondents. The sampling procedure used by the U.S. Census Bureau (2009) and Williams (2008) is outlined next using the notation described in Section 2.1. We introduce here the sampling fraction, r_s, where s specifies the state. A sampling fraction of r_s signifies that roughly 1 in $1/r_s$ GQ persons are sampled in a year from that state.

[3] This model is based on 2006-2009 ACS data. Not much information exists on the behavior of small GQs across time; as such, we do not model population changes in those GQs in years $y > 1$. See Joyce (2010) for more details.

ACS Sample Simulation Algorithm

1. Order the 2000 GQ universe list by GQ type, county, tract, block, and finally the unique identification GQ number. (Note: All GQs from the same state have the same region and division codes.)
2. Assign each GQ i to the small stratum if $o_{i,k,c,t,0} \leq 15$ and to the large stratum if $o_{i,k,c,t,0} > 15$.
3. Assign each small GQ (ordered as in Step 1) to one of five subframes. The first small GQ is assigned to subframe 1, the second to subframe 2, and so forth; the sixth small GQ is then assigned to subframe 1. In any particular year, samples are chosen from only one subframe; in any collection of five years, no subframe is repeated. For our purposes, it is not necessary to assign new GQs to a subframe because (a) we are not adding new GQs to the universe and (b) large GQs which become small GQs during the course of the simulation cannot be sampled again in our five year time period.
4. For each year,

   (a) If $y = 1$, let the expected population for each GQ be $e_{i,k,c,t,1} = o_{i,k,c,t,0}$. For subsequent years, $e_{i,k,c,t,y} = o_{i,k,c,t,y'}$ where $y' < y$ and indicates the most recent year for which a current population has been *observed*. Note that $y' = 0$ *only* if the GQ has never been sampled before. (See Step 4h for details.)

   (b) Draw a systematic sample (every $1/(5r_s)$th observation) of small GQs from the designated subframe[4]. For year 1, draw from subframe 1, from year 2, draw from subframe 2, and so forth. In year 6, draw from subframe 1 again. If there are $n > 1/(5r_s)$ GQs in a subframe:

      i. Choose a random starting point, S_0, from $Unif(1, 1/(5r_s))$, where $Unif(a, b)$ is the continuous uniform distribution from a to b with the endpoints included.

      ii. Construct a vector by continuously adding $1/(5r_s)$ to S_0: S_0, $S_1 = S_0 + 1/(5r_s)$, $S_2 = S_0 + 2/(5r_s)$, …, $S_L = S_0 + L/(5r_s)$, where L is the largest integer such that $S_L \leq n$.

      iii. If we number the GQs from 1 to n, ordered as in Step 1, the GQs selected to be sampled are given by $\lceil S_l \rceil$ where $l = 0, \ldots, L$ and $\lceil \cdot \rceil$ is the ceiling function.

   If, however, $n \leq 1/(5r_s)$, then select one GQ at random for sampling.

   (c) For large GQs,

      i. Calculate the measure of size, m_i, for each large GQ i by dividing the total number of persons expected in that GQ by 10: $e_{i,k,c,t,y}/10$. Next, calculate the cumulative measure of size for each large GQ i, by computing $\sum_{j=1}^{i} m_j$ (maintaining the ordering in Step 1). Let M be the maximum cumulative measure of size; that is, the sum of all of the measures of size. To illustrate these intermediate steps, we provide a simple example in Table 3. The first column is the GQ identifier and the second column lists the expected population size of each GQ. The third column provides the measure of size and the final column, the cumulative measure of size. For these example data, $M = 140.5$.

      ii. To obtain a systematic sample of 1 in $1/r_s$ persons, the cumulative measure of size is used as follows:

         A. Draw a random starting point, S_0, from $Unif(0, 1/r_s)$.

[4] We draw every $1/(5r_s)$th GQ instead of $1/r_s$th GQ because we split the list of GQs into 5 subframes.

Table 3: Example of Calculating the Cumulative Measures of Size

GQ (i)	$e_{i,k,c,t,y}$	m_i	$\sum_{j=1}^{i} m_j$
1	55	5.5	5.5
2	1000	100	105.5
3	20	2	107.5
4	30	3	110.5
5	300	30	140.5

B. Create a vector of cumulative measure of size values by continuously adding $1/r_s$ to S_0. That is, $S_0, S_1 = S_0 + 1/r_s, S_2 = S_0 + 2/r_s, \ldots, S_L = S_0 + L/r_s$, where L is the largest integer for which $S_L \leq M$ is true. Denote this vector as $\{S\}$. To continue with the example in Table 3, let $r_s = 0.025$ (so $1/r_s = 40$) and $S_0 = 3.7$. Then, $\{S\} = \{3.7, 43.7, 83.7, 123.7\}$ and $L = 3$.

C. For each value of S_l, the GQ with the cumulative measure of size that is the supremum of S_l is selected to be in the sample. Repeat for all elements in $\{S\} = \{S_0, S_1, \ldots, S_L\}$. When a GQ is selected for the sample, it is called a hit. Multiple hits from the same GQ are possible if there are a large number of persons living there. The selected GQs from the example in Table 3 are 1, 2, 2, and 4 given $\{S\}$. Note that GQ 2 is hit twice in this example; therefore, two sets of 10 persons are sampled from this GQ.

(d) All persons in a chosen GQ hit are assigned to the same month of data collection. A random starting point is chosen from 1 to 12, say p. Keep the order of GQ hits as is from the GQ sampling steps and append the sampled large GQs to the sampled small GQ list. From the collection of chosen GQs, the first GQ hit is assigned to be sampled in the month given by index p in Table 4. For example, if $p = 2$, then these persons are surveyed in July. The second GQ hit is assigned to the month corresponding to $p + 1$ and so forth[5]. After $p = 12$ is reached, the index returns to $p = 1$ (Williams, 2008).

Table 4: Survey Month Assignment (Williams, 2008)

Index (p)	Month
1	March
2	July
3	November
4	February
5	June
6	October
7	January
8	April
9	November
10	May
11	August
12	December

(e) To incorporate seasonality, it is assumed that college dormitories are empty during the summer months of June, July, and August. If such a GQ is assigned to be sampled during

[5] Some GQ types are assigned to sample months differently; however, to reduce complexity and because they will have a negligible affect on the results, we do not apply those rules here.

any of these months, then zero persons are selected for those GQs.

(f) Some GQs, once selected to be in the sample, are found to not exist. No person records are selected for such GQs.

(g) For those GQs which are sampled (and exist), with the exception of college dormitories sampled in summer months, the next step is to sample person records. The samples are taken from the population simulated for that year.

  i. For small GQs,

    A. If $o_{i,k,c,t,y} \leq 15$: select all persons in the GQ to be included in the sample.

    B. If $o_{i,k,c,t,y} > 15$: select 10 persons without replacement to be included in the sample. Reassign this GQ to the large stratum for the following year.

  ii. For large GQs, for each hit,

    A. If $o_{i,k,c,t,y} > 15$: select 10 persons without replacement to be in the sample.

    B. If $10 < o_{i,k,c,t,y} \leq 15$: select 10 persons without replacement to be in the sample. Reassign this GQ to the small stratum for the following year.

    C. If $o_{i,k,c,t,y} \leq 10$: select all persons to be in the sample. Reassign this GQ to the small stratum for the following year.

    D. If the same GQ is hit $n > 1$ times in the same month, then $10 \times n$ persons must be drawn without replacement from the GQ population to be in the sample. If, by chance, $10 \times n > o_{i,k,c,t,y}$, then choose all persons in the GQ to be in the sample.

(h) Replace the expected population for the next year, $e_{i,k,c,t,y+1}$, with the current population value, $o_{i,k,c,t,y}$, for all sampled GQs [6]. The exception to this rule is for college dormitories. In this case, only apply the updates for GQ hits which have sample dates outside of the summer months. In addition, reclassify each sampled GQ i to the small stratum if $o_{i,k,c,t,y} \leq 15$ and to the large stratum if $o_{i,k,c,t,y} > 15$.

In the simulation, as the population size of each large GQ is updated every year, it is possible for a large GQ to become a small GQ. However, these changes are only detected when a GQ is sampled and we ascertain the true population size. To simplify the simulation, for such cases we do not add the GQ to the small stratum subframe and therefore do not allow it be sampled again in the 5 year simulation period. Finally, once a GQ becomes a small GQ, the population size remains fixed for the remainder of the years. Note that college dormitories can only switch strata and expected populations if they have hits outside the summer months.

3 Imputation Methodology

Given the complexity of the group quarters population, the procedure for sampling this population, and the level of coverage that is desired, careful consideration must be given to both the selection of GQ facilities for imputation and to the selection of donor records. As such, we consider two methods for the selection of group quarters facilities, and two methods for the selection of donors. The imputed records are appended to the sampled records to form a complete augmented data set.

Both GQ selection procedures select all large-stratum facilities for imputation. However, one procedure first selects facilities needed to produce non-zero 1-, 3- and 5-year county-level estimates, and then selects small-stratum facilities required to produce 5-year tract-level estimates; the second procedure reverses this order. In both GQ selection methods, the fraction of residents to be imputed

[6] GQs chosen to be sampled for the following year are actually chosen before data collection for the current year is complete. We do not incorporate this into our procedure because it has a negligible effect on the results.

is chosen not only to resemble ACS sampling rates but also to produce reasonable variance estimates. The details for both procedures are given in Section 3.1[7].

Once facilities have been selected for imputation, we choose donor records to populate the facilities according to one of two methods. Both donor selection procedures give preference to donors from within the same specific GQ type, and ensure that donors come from within the same major GQ type as the recipient GQ. However, one method focuses on finding donors from facilities which are geographically close while the other focuses on finding donors from facilities in geographies which are demographically similar. The details of these procedures are given in Sections 3.2 and 3.3.

3.1 Selection of GQ Facilities

The first GQ selection procedure gives priority to obtaining representation for each major GQ type group in each county. Then facilities are selected to establish representation for each major type group at the tract level. A detailed outline of the procedure is given next. We use the notation provided in Table 2 here again. Let $\lfloor \cdot \rceil$ denote the nearest integer function.

GQ Selection Procedure 1: County-Level Coverage First

1. For each year y and each large GQ not in sample i, impute $\max(1, \lfloor 0.025 \times e_{i,k,c,t,y} \rceil)$ records.
2. For each year y and for each combination of county c and major GQ type k on the year's frame that is not in the year's sample nor in the year's imputes, select a small GQ facility at random with probability equal to the reciprocal of the number of small GQ facilities in the county and of the same major GQ type.
3. For each GQ i selected in Step 2, impute $\max(1, \lfloor 0.2 \times e_{i,k,c,t,y} \rceil)$ records.
4. Select all combinations of tract and major GQ type that exist on any year's sampling frame but are not in any year's sample, nor in any year's imputed records.
5. For each combination of tract t and major GQ type k in Step 4, for each year y that the combination exists on the sampling frame, select a small GQ facility at random with probability equal to the reciprocal of the number of small GQ facilities in the tract and of the same major GQ type.
6. For each GQ i selected in Step 5, impute $\max(1, \lfloor 0.2 \times e_{i,k,c,t,y} \rceil)$ records.

The second GQ selection procedure gives priority to obtaining representation for each major GQ type group in tracts that are not sampled in the 5-year period. Then facilities are selected to produce non-zero 1- and 3-year estimates for each major type group and county combination. Details of the procedure are as follows.

GQ Selection Procedure 2: Tract-Level Coverage First

1. For each year y and each large GQ i not in sample, impute $\max(1, \lfloor 0.025 \times e_{i,k,c,t,y} \rceil)$ records.
2. Select all combinations of tract and major GQ type that exist on any year's sampling frame and are not in any year's sample, nor in any year's imputed records.

[7] The procedures for selecting facilities for imputation were developed through consultation with the group quarters small area estimation group at the U.S. Census Bureau.

3. For each combination of tract t and major GQ type k in Step 2, for each year y that the combination exists on the sampling frame, select a small GQ facility at random with probability equal to the reciprocal of the number of small GQ facilities in the tract and of the same major GQ type.

4. For each GQ i selected in Step 3, impute $\max\left(1, \lfloor 0.2 \times e_{i,k,c,t,y} \rceil\right)$ records.

5. Select all combinations of county c and major GQ type k which are on the year 5 sampling frame but not in the year 5 sample, nor in the year 5 imputed records.

6. For each combination of county c and major GQ type k in Step 5 and for each year in the range 3-5 that the combination exists on the sampling frame and is not in the year's sample nor in the year's imputed records, select a small GQ facility at random with probability equal to the reciprocal of the number of small GQ facilities in the county and of the same major GQ type.

7. For each GQ i selected in Step 6, impute $e_{i,k,c,t,y}$ records.

3.2 Imputation by Geography

The first donor selection procedure chooses from within specific type when the donor to imputation ratio within specific type is reasonable, and gives gives preference to donors from facilities that are geographically close. Once GQ facilities have been selected for imputation, the donor pool for each facility is set to be the first combination of geography and GQ type in the following list in which there is at least one donor per five imputed records needed: (1) county and specific GQ type, (2) county and major GQ type group, (3) state and specific GQ type, (4) state and major GQ type group, (5) division and specific GQ type, (6) division and major GQ type group, (7) region and specific GQ type, (8) region and major GQ type group, (9) nation and specific GQ type, and (10) nation and major GQ type group. For example, suppose that in a particular county we wish to impute one hundred records into college dormitories. If at least twenty dormitory residents in the county have been interviewed, we sample these interviews for imputation, with replacement, one hundred times. If fewer than twenty dormitory residents in the county have been interviewed, we expand the geography of the donor pool (to the state, division, region, or nation) as necessary so that there are at least twenty records from which to sample. Census divisions and regions are summarized in Appendix A, Table 13.

3.3 Imputation by Cluster

As an alternative to a "nearest neighbor" imputation method like that of Section 3.2, we consider application of the K-means clustering algorithm to find donors from facilities located in tracts that are demographically similar. The Tract-Level Planning Database (PDB) is a collection of household, demographic, and socioeconomic variables assembled to identify the reasons why people are missed in the Census (Bruce and Robinson, 2007). To help design a targeted marketing campaign for the 2010 Census, Bates and Mulry (2008) perform a K-means cluster analysis applied to census tracts on the following twelve PDB variables:

1. Percent of housing units which are vacant,

2. Percent of housing units that are not single detached or attached,

3. Percent of housing units occupied by renters,

4. Percent of occupied housing units with more than 1.5 persons per room,

5. Percent of households which are not husband/wife families,
6. Percent of occupied housing units with no telephone service,
7. Percent of persons who are not high school graduates (ages 25+),
8. Percent of people below poverty,
9. Percent households with public assistance income,
10. Percent of unemployed people,
11. Percent of linguistically isolated households[8], and
12. Percent of occupied units where householder moved into unit in the last calendar year.

Additional PDB variables are highly correlated with these twelve, do not add to the analysis, and are therefore excluded. Analysis using the above variables revealed the following eight distinct clusters of tracts:

1. **All Around Average I (homeowner skewed)**. Bates and Mulry (2008) describe this group as the "average Joe." They are close to the national average on nearly all 12 PDB variables used in the analysis.
2. **All Around Average II (renter skewed)**. This group is similar to the All Around Average I group, but has a higher percentage of renters and multi-unit residences.
3. **Economically Disadvantaged I (homeowner skewed)**. As the name indicates, this group is comprised mostly of economically disadvantaged individuals. It is similar to the following cluster, but more than half of this group are homeowners.
4. **Economically Disadvantaged II (renter skewed)**. This cluster has an overwhelmingly high percentage of renters (84%). It contains the highest percentage of people in poverty, unemployed, and on public assistance. Fifty-four percent of this group is black, and the number of Hispanics in this cluster is higher than average.
5. **Ethnic Enclave I (homeowner skewed)**. This group is predominantly Hispanic (61%), has above-average percentage of persons in poverty, on public assistance, percentage of crowded units, and percentage of adults (25+) without a high school diploma.
6. **Ethnic Enclave II (renter skewed)**. This cluster is also predominantly Hispanic (59%), is exclusively urban, densely populated, has a high percentage of linguistically isolated households, and less than half of the individuals in these tracts have a high school diploma. On average, 75% of this group are renters.
7. **Single/Unattached/Mobiles**. This group is comprised mostly of young, single renters with higher-than-average education and high mobility. 59% of this group is non-Hispanic white, followed by black at 17%, and this cluster has a higher-than-average percentage of Asian residents (7%).
8. **Advantaged Homeowners**. This cluster is the least densely-populated and has the lowest levels of unemployment, poverty, percentage of renters, and non-spousal households. This group has little racial diversity and is 85% non-Hispanic white.

[8] A linguistically isolated household is one in which a language other than English is spoken and no person (ages 14+) speaks English very well.

For a more detailed description of these clusters, please see Bates and Mulry (2008).

We use the clusters to guide donor selection in the following manner. Once GQ facilities have been selected for imputation, we first group facilities by cluster and specific GQ type. For each combination of cluster and specific GQ type, if there is at least one donor per five imputations needed, donors are selected at random from within cluster and specific type.

If the donor to imputation ratio is less than 0.2 and a facility is in clusters 1-6, clusters 1 and 2 are combined, clusters 3 and 4 are combined, and clusters 5 and 6 are combined, resulting in the following five clusters: (1) All around Average I & II, (2) Economically Disadvantaged I & II, (3) Ethnic Enclave I & II, (4) Single/Unattached/Mobiles, and (5) Advantaged Homeowners. If the new donor to imputation ratio is at least 0.2, donors are selected from within specific type and new cluster. If the donor to imputation ratio is less than 0.2, donors are selected at random from within specific GQ type. In very few cases (a fraction of a percent, on average), the donor pool is expanded to major GQ type group. A summary of the number of donors selected from within the original eight clusters, the collapsed clusters, specific GQ type and major GQ type group is given in Section 4.

3.4 Weighting

Asiala (2010) has developed a weighting scheme to accompany any of the four proposed imputation methods and is applied to the augmented data: the data set containing both the sampled and imputed records. This procedure ensures that when computing estimates for small areas, the weights only represent persons within that tract or county. We briefly describe this procedure next.

The base weights for each year are assigned in one of two ways. Observations in GQs which have an expected or observed count above 15 are weighted so that they represent persons in that GQ only. This is because, under any imputation method, all "large" GQs are represented. That is, each GQ in this category has either sampled or imputed residents. On the other hand, not all "small" GQs, which house 15 or fewer residents, are represented in the augmented data. Therefore, these GQs require an alternate approach. Observations in GQs where the expected or observed population is 15 or fewer are weighted so that they represent residents of all GQs with an expected or observed population of 15 or fewer in that tract. For multi-year estimates, the base weights are simply divided by the period length (3 or 5).

The second step is to apply the population controls which fix the total population at the state level by major type (1-7). These are applied through tract and county constraints in order to control population totals at these lower levels of geography from the outset.

4 Results

A simulation study such as ours offers two distinct advantages: (a) we know the "truth" because we have simulated the population and (b) we can directly compute measures such as bias and mean squared error since we have drawn multiple, independent samples. The results for 1-, 3-, and 5-years are similar and our discussion covers all three periods. We present the main results in this section. A complete description of the results can be found in Nagaraja (2010a,b,c,d,e). All graphs referred to in this section can be found in Appendix B.

To evaluate the results of the imputation methods, we analyze the estimates from the five methods: one design-based method and four imputation-based methods. These are summarized below.

1. Design-based: estimates are computed using the sampled data only (denoted as "sample" in plots).

2. Expanding Search-County Imputation: expanding search, county-level coverage handled first (denoted as "expand county").

3. Expanding Search-Tract Imputation: Expanding search, tract-level coverage handled first (denoted as “expand tract”).

4. K-means Search-County Imputation: K-means search, county-level coverage handled first (denoted as “k-means county”).

5. K-means Search-Tract Imputation: K-means search, tract-level coverage handled first (denoted as “k-means tract”).

The Census 2000 short form data can be used to produce counts of the following basic demographic variables:

1. total population,

2. sex: male/female,

3. age: 0-17/18-34/35-64/65 years and older,

4. ethnicity: Hispanic/not Hispanic, and

5. race: white, not Hispanic/black, not Hispanic/other, not Hispanic.

For each characteristic[9], geography, and major type group, we have five sets of estimates and a true, population value. Let n be the total number of samples ($n = 25$ here). Then, let $x_{v,m,k,g}$ be the true (population) value of variable v, period m, for major GQ type k, and in a given geography g. Let, $\hat{x}_{v,m,k,g,l,s}$ be the estimated value of a variable v for geography g, in period m, GQ type k, using method l, and sample s. Define $\bar{\hat{x}}_{v,m,k,g,l}$ as $\sum_{s=1}^{n} \hat{x}_{v,m,k,g,l,s}/n$. Finally, let $l = 0$ represent the design-based, current method (sampled records only), and $l = 1, \ldots, 4$ represent the four proposed methods. An estimate is computed by adding the final weights of those records in the geography, period, and sample which have the particular characteristic v. This notation is summarized in Table 5.

Table 5: Additional Notation for Evaluation

Symbol	Definition
n	total number of samples (25)
v	variable (characteristic)
k	major GQ type (1-7)
m	period for 1-, 3-, or 5-year estimates
g	geography: a specific tract/county/state
l	estimation method (0 for design-based, 1-4 for imputation-based)
p	sample (one sample is set of 5-years)
$x_{v,m,k,g}$	true (population) value of variable v, in period m, of GQ type k, and for geography g
$\hat{x}_{v,m,k,g,l,p}$	estimated value of variable v, in period m, of GQ type k, for geography g, using method l, and for sample p
$\bar{\hat{x}}_{v,m,k,g,l}$	average estimated value of variable v, in period m, of GQ type k, for geography g, using method l over all n samples ($\frac{1}{n}\sum_{p=1}^{n} \hat{x}_{v,m,k,g,l,p}$)

[9] Results for the variable “total population” are not evaluated at the state level because these are set to the population controls and therefore will always, by construction, be correct.

4.1 Donor Analysis

The first step in evaluating the imputation-based methods is to determine how many imputed records the proposed schemes generate. Overall, the number of sampled records is comparable to the number of imputed records. However, if we categorize records by major GQ type, we see very different patterns as shown in Table 6. In this table, the sampled and imputed counts from year 5 in sample 1 (out of 25) are listed for each GQ type at the national level for the "expand county" method[10]. The bottom row provides the total counts if all major GQ types are aggregated. Not only are the number of imputed records highly variable across GQ types, the ratio of sampled to imputed records varies considerably as well. Two reasons for this result are (a) the number of persons in the nation residing in each of the GQ types differ and (b) the average size of the GQs varies by GQ type.

Table 6: Number of Sampled and Imputed Records (sample 1, year 5)

GQ type	Num. records sampled	Num. records imputed
Correctional institution (1)	52,916	16,585
Juvenile detention (2)	1,623	4,695
Nursing home (3)	27,536	30,871
Other long-term care (4)	1,877	5,799
College dormitory (5)	33,280	26,798
Military facility (6)	5,404	3,645
Other non-institutional (7)	10,243	41,359
Total	132,879	129,752

These patterns can be further understood by examining Figures 2 and 3. For this pair of plots, the average number of sampled and imputed records by year for GQ types 6 (military facilities) and 2 (juvenile detention facilities) across samples is graphed. The x-axis represents year and, within a year, each bar represents one of the five methods. The y-axis shows the average number of records sampled or imputed across samples. The variation of counts across samples was very small and so was omitted from the graph.

There are two features to note here. First, in general, the number of sampled records decreases as year increases. This attribute is simply an artifact of allowing GQs to close. The delete (or closing) rates are high for some GQ types, making this pattern even more pronounced. In particular, large GQs close at a much lower rate than do small ones. GQ types containing larger GQs therefore have smaller changes from year to year in the number of records sampled which leads to smaller changes in the number of records imputed from year to year. We can see these types of patterns in both Figure 2 and Figure 3. The number of imputed records, on the other hand, does not vary significantly across years nor across imputation methods. This pattern holds for all GQ types and occurs because the same level of data coverage is desired irrespective of the imputation methodology.

The second feature is whether the number of imputed records is greater or fewer than the number of sampled records. There are some clear patterns based on GQ type:

1. number imputed $<$ number sampled: GQ types 1 (correctional institutions), 5 (college dormitories), and 6 (military facilities) (see Figure 2)

2. number imputed $>$ number sampled: GQ types 2 (juvenile detention facilities), 4 (other long-term care facilities), and 7 (other non-institutional facilities) (see Figure 3)

3. number imputed $\approx$ number sampled: GQ type 3 (nursing homes)

[10] We obtain similar results for the "expand tract" method.

The relationships between the number of records imputed and the number sampled are a result of individual GQs varying considerably in size. GQ types 1, 5, and 6 tend to be very large, and larger GQs are selected to be in sample more often than smaller GQs. Therefore, the GQs representing a large percentage of the residents are selected for sample and therefore do not require imputation to obtain county and tract coverage. GQs of types 2, 3, 4, or 7 tend to be much smaller, requiring more imputed records. If we examine the ratio of imputed to sampled records broken down by state, the same patterns hold in general.

The next step is to determine where donors are found in relation to the imputed record. Our goal is to find donors which are the "closest match" or "most similar" to persons who live in the imputed GQ. Defining the appropriate measure of "most similar" is the difficult part. We proposed in Section 3 two methods of donor selection: an expanding search and a K-means clustering. The former method defines a close match by geography and GQ type and the latter matches based on GQ type and characteristics of the surrounding household population, without regard to geographic proximity.

Table 7: Geographic Relationship of Donor Record to Imputed Record

GQ Type	Average Num. Residents*	Expanding Search	k-means Search
Correctional institution (1)	2,501,305	within county	outside state
Juvenile detention (2)	93,374	within state	outside state
Nursing home (3)	1,474,189	within county	outside state
Other long-term care (4)	133,579	within state	outside state
College dormitory (5)	2,265,390	within county	outside state
Military facility (6)	451,095	within state	outside state
Other non-institutional (7)	843,973	within county	outside state

* national average across 5 years in the simulated population

Regardless of which method is applied, most donors are fortunately of the same specific type[11]. What differs among methods is the geographical source of the donor – these results are summarized in Table 7. For the expanding search methods, most donors are found within the state and for some GQ types within the county of the imputed GQ. Major GQ types with more residents (and therefore more sampled residents) tend to have donors found within the same county as the imputed GQ. For those GQ types with fewer residents, donors are most often found outside of the county but within the state. For the K-means methods, however, nearly all donors are found outside of the state. This is due to the distribution of the clusters across the nation (given in Appendix A, Table 14). For instance, 65% of tracts belonging to the "Ethnic Enclave II, Renter Skewed" cluster are located in California and New York. Furthermore, the distributions of GQs across the nation given major GQ type are heavily skewed. As an illustration, consider the distribution of correctional institutions in the Ethnic Enclave II cluster (6), given in Table 8.

Fifty-seven percent of all correctional institutions located in tracts belonging to the Ethnic Enclave II cluster are in California. Therefore, any of the other thirteen states with correctional institutions in tracts in this cluster (which is predominantly Hispanic, urban, densely-populated, and linguistically isolated) and in need of imputation are likely to receive donors from California. Analogous stories can be told for other combinations of major GQ type group and cluster.

We also examine the number of times each sampled record is used for imputation. Recall that, for each method, at each step of the imputation, we restrict the donor pool to contain at least one donor per five imputed records needed but we do not impose a strict ceiling on the number of times a record can serve as a donor. Ideally, we want any single donor to be used few times; if donors are used many times, the estimates are more likely to be biased and the variance misleading. For

[11] GQ types 3 and 5 have only one specific type.

Table 8: Geographical Distribution of Correctional Institutions in the Ethnic Enclave II Cluster

State	No. Correctional Institutions	Proportion
Arizona	7	0.04
California	98	0.57
District of Columbia	2	0.01
Florida	21	0.12
Georgia	3	0.02
Hawaii	1	0.01
Illinois	1	0.01
Massachusetts	16	0.09
Nevada	5	0.03
New York	4	0.02
Rhode Island	1	0.01
Texas	4	0.02
Washington	1	0.01
Wisconsin	7	0.04
Total	171	1.00

the K-means selection procedure, 99% of sampled records serve as donors seven or fewer times and 95% of sampled records are used as donors four or fewer times. For the expanding search method, 99% of sampled records serve as donors eight or fewer times and 95% of sampled records are used as donors five or fewer times. The maximum number of times a donor is selected for imputation is twenty, on average across samples, for both donor selection procedures.

4.2 Summary Measures

We are able to directly compute the following measures because multiple independent samples are drawn from the simulated 5-year population:

$$Bias_{v,m,k,g,l} = \bar{\hat{x}}_{v,m,k,g,l} - x_{v,m,k,g} \tag{1}$$

$$Variance_{v,m,k,g,l} = \frac{1}{n-1}\sum_{p=1}^{n}\left(\hat{x}_{v,m,k,g,l,p} - \bar{\hat{x}}_{v,m,k,g,l}\right)^2 \tag{2}$$

$$MSE_{v,m,k,g,l} = \frac{1}{n}\sum_{p=1}^{n}\left(\hat{x}_{v,m,k,g,l,p} - x_{v,m,k,g}\right)^2 \tag{3}$$

$$MAD_{v,m,k,g,l} = \frac{1}{n}\sum_{p=1}^{n}\left|\hat{x}_{v,m,k,g,l,p} - x_{v,m,k,g}\right| \tag{4}$$

$$\begin{aligned} \%MAD_{v,m,k,g,l} &= \frac{\frac{1}{n}\sum_{p=1}^{n}\left|\hat{x}_{v,m,k,g,l,p} - x_{v,m,k,g}\right|}{x_{v,m,k,g}} \times 100 \\ &= \frac{MAD_{v,m,k,g,l}}{x_{v,m,k,g}} \times 100 \end{aligned} \tag{5}$$

Each of these measures allow us to compare the performance of the five methods (four imputation-based method and one design-based method) across the 25 samples.

We find that the imputation-based methods are all more biased than the design-based method at each geography level: tract, county, state. In Figure 4, the bias at the tract, county, and state levels is plotted for the 5-year estimates of the number of persons who are white but not Hispanic. The box plots can be interpreted as follows. On the x-axis, each group of box plots represents a GQ major type. The y-axis indicates the bias. Each individual box plot within a group represents a method and represents the range of bias values across all tracts/counties/states for the demographic variable and major GQ type combination. In each plot, we see that the range of bias is larger for the imputation-based methods than for the design-based method (in gray) for all levels of geography.

A second, more specific, example can be seen in Figure 5. The absolute bias[12], variance, root mean squared error, and mean absolute deviation are plotted for the 5-year estimates of the number of females who live in a correctional facility (GQ type 1) in a county. For each of these plots, the estimate using the "expand county" method of a measure is plotted on the x-axis. The design-based values of that measure are plotted on the y-axis. Each point represents a specific state in the U.S. or the District of Columbia.

To interpret these plots more easily, we add 3 features. First, the 45° line is plotted (dotted line). If a point is above the 45° line, then the design-based method has a higher value of the measure indicating the imputation-based method performs better in the geography represented by that point; if a point is below the 45° line, then the converse is true. On each plot, the total number of states which are above and below the 45° line are printed (points directly on the line are excluded here). Finally, because many points may overlap, areas which contain more points are shown in a darker color than those with fewer points.

Consider the plot for absolute bias (on the top-left). A lower value of absolute bias is more desirable, regardless of the method applied. Therefore, knowing which method (between the design-based and the "expand county" methods) has a lower value of a measure is informative. From this plot we can see that most of the points are below the 45° line indicating that the "expand county" method produces more biased estimates. Upon further investigation, we find that the imputation procedures are imputing more males into correctional facilities than females. As a result, we systematically undercount the number of females (negative bias) and systematically over count the number of males (positive bias) in correctional facilities. One way to mitigate this problem would be to identify which correctional facilities are male-only, female-only, or mixed and find donors accordingly.

If we refer back to Figure 5, we see that the variances (top-right plot) are generally smaller for the imputation-based estimates when compared to the design-based estimates. That is, most of the points on the graph are above the 45° line. This is the case across all combinations of variable, GQ type, period length, and imputation method. The results, however, are mixed for the MAD and MSE analyses. We can also see this in Figure 5 (bottom two plots) because the points are scattered randomly around the 45° line. For both of these measures, the imputation-based methods tend to perform better across counties for variables which "behave" like total population such as gender, white (not Hispanic), and total population. For the MSEs, the design-based estimates tend to perform better across both counties and tracts for variables which measure "subpopulations" such as the individual age groups and race other (not Hispanic). These relationships do not always hold for either the MAD or MSE at the state level, however.

In the set of plots in Figure 6, we graph bias, standard deviation, MSE, MAD, and % MAD for 5-year "expand tract" estimates at the county level for different variables and GQ types. These measures are plotted against either the total number of persons or the true value of the variable. Each point on the plot represents a particular county. For plots of %MAD, note that the measure is undefined if the true value is 0 and is an indeterminate form if both the MAD and the true value

[12] Bias is plotted as absolute value of bias, or $|bias|$. This is because bias can be any real number which causes problems when comparing values. For instance, if the bias for the design-based method is 10 and the bias for the imputation-based method is -20, doing a direct comparison would indicate that since $-20 < 10$, the imputation-based method has a lower bias. However, this is not the case. $|-20| > |10|$ and therefore the imputation-based method has a higher bias although in the opposite direction. In this analysis, we only care about magnitude not direction because bias in any direction is undesirable.

are 0. Such cases are omitted from the plot; however, the frequency of each case is printed on the plot.

4.3 Seasonality

In an effort to incorporate seasonal variation into the simulated population, we decided that college dormitories sampled during the summer have no residents. However, such GQs are candidates for imputation in our proposed methods. If a GQ is large (more than 15 persons) and has no samples within a year, then it will be automatically imputed into. On the other hand, if a county or tract has no representation (no sampled or imputed records), it may be necessary to impute such a GQ. These imputations will inflate the number of GQ residents throughout the year because the seasonal component is ignored.

The first step in our analysis is to determine how often either of these situations occur. Therefore, we count the number of GQs which are sampled in the summer and have no residents in the following five categories within a specific year:

1. GQs with more than 15 residents ($o_{i,k,c,t,y} > 15$) which are also sampled at an alternate time of the year (multiple hits),

2. GQs which are thought to have more than 15 residents ($e_{i,k,c,t,y} > 15$) which are imputed into,

3. GQs with fewer than 15 residents ($o_{i,k,c,t,y} \leq 15$) which are also sampled at an alternate time of the year (these were expected to have more than 15 residents and were assigned multiple hits–this case is very rare),

4. GQs which are thought to have 15 or fewer residents ($e_{i,k,c,t,y} \leq 15$) and are imputed, and

5. GQs which are thought to have 15 or fewer residents ($e_{i,k,c,t,y} \leq 15$) and are neither imputed nor sampled at another time.

Combining categories 1 and 2 gives us the total number of college dormitories sampled during the summer with an observed or expected population greater than 15 residents. Combining categories 3 through 5 give us the total number of college dormitories sampled during the summer with an observed or expected population of 15 or fewer.

In Table 9, the mean number of GQs in each category across samples for each year is shown. Standard deviations are given in parentheses below. The results are for the expanding-county method; however, we do not expect the results to change substantially across methods. This is because the first step in all of the imputation-based methods is to impute all GQs not in sample which are expected to have more than 15 residents and most college dormitories fall into that category. We see that for case 2, we have roughly 1,000 persons imputed each year. This is not an insignificant amount; therefore, we may want to limit imputation for such cases to improve our estimates.

4.4 Geographies with No Imputed Records or No Sampled Records

Assume a state has two counties A and B, only one GQ type, and GQ persons living in both counties. In a given year, assume only people from county A are sampled. If we apply the imputation procedure to this setting, we would then impute people only into county B to obtain county representation.

As a consequence of the design-based weighting scheme, the estimated number of people living in county A would be the sum of the weights assigned to each sampled person from county A. To obtain the county total estimate for county B, we would do the same; as there were no people sampled in county B, the estimated total is 0. This means that the estimate for county A is too high and the estimate for county B is too low. The design-based weighting procedure controls population totals only at the state level, not at lower geographies. Therefore, sampled persons in county A are

Table 9: College Dormitory Sampling and Imputation Rates

	Category	Year 1	Year 2	Year 3	Year 4	Year 5
1	$o_{i,k,c,t,y} > 15$, sampled	157.36	159.32	164.00	182.24	203.16
		(7.22)	(8.38)	(10.85)	(11.10)	(10.45)
2	$e_{i,k,c,t,y} > 15$, imputed	1,088.12	1,033.40	988.40	943.44	904.64
		(11.34)	(11.25)	(13.23)	(13.17)	(13.38)
3	$o_{i,k,c,t,y} \leq 15$, sampled	2.88	1.00	0.84	1.24	0.88
		(1.01)	(0.81)	(0.85)	(1.23)	(0.72)
4	$e_{i,k,c,t,y} \leq 15$, imputed	1.96	1.64	2.32	2.08	0.76
		(1.27)	(1.22)	(1.06)	(1.38)	(0.87)
5	$e_{i,k,c,t,y} \leq 15$, neither	34.00	28.40	24.76	20.72	18.12
		(3.40)	(3.97)	(4.24)	(4.20)	(3.39)

weighted so that they represent persons in county B resulting in incorrect estimates for both county A and B by construction.

On the other hand, the imputation-based weighting scheme controls the population totals at the tract-level for 5-year estimates and at the county-level for the 1- and 3-year estimates preventing this type of situation from occurring. This weighting scheme allows the sampled cases from county A to represent only persons living in county A. Similarly, the imputed cases in county B are weighted to represent only persons who live in county B[13]. This is a highly desirable property of the proposed methodology and should improve the design-based estimates.

Tracts or counties in which all of the records are imputed are geographies for which there are no sampled records. Such a situation allows us to isolate the imputed estimates because there is no contribution from sampled data. Therefore, the design-based estimate for any variable is always zero in such a case. If we count the total number of persons, this estimator is always negatively biased if there are, in fact, people living in the county. For specific characteristics, the design-based estimates are negatively biased only if people exhibiting that characteristic reside in the county.

In geographies for which there is no sample, the imputation-based methods could improve estimates. One example is shown in the top two plots of Figure 7 for 3-year estimates of the number of people between the age of 18 and 34 in each county in one sample. In Table 10, the number of counties which have no sampled records by GQ type are provided to illustrate how many points are used to construct each box plot.

To construct these plots, we compute the errors (estimate - true value) for the total number of people in each county with no sampled records for one sample. These are plotted for each county and for each method and GQ type combination, in the form of box plots. There is an improvement for GQ types 1 (correctional facilities), 5 (college dormitories), and 6 (military facilities). These are the GQ types for which we expect a fair number of people in this age group. Because the method for selecting GQs to impute into is the same for the pair “expand county” and “k-means county” and for the pair “expand tract” and “k-means tract”, we plot these pairs together, respectively.

Sometimes, we see a reduction in the errors for some variables and GQ types when comparing the imputation methods with the design-based method. This improvement is most likely an effect of the the imputation-based population constraints at the tract and county levels. The bottom two plots in Figure 7, constructed in the same way as the first, are examples of this feature. The errors are smaller for GQ major types 2 (juvenile facilities) and 5 (college dormitories). As before, in Table 10, the number of counties which had no imputed records are listed by GQ type.

[13] Note that sampled County A persons are indirectly influencing estimates in county B because they act as donors; however, the imputation-based weighting scheme apportions the weights in such a way to separate out the weights by county.

Table 10: Number of Counties with No Imputed or No Sampled Records (3-year Estimates)

GQ type	All imputed	All sampled
Correctional institution (1)	1,134	87
Juvenile detention (2)	781	5
Nursing home (3)	836	8
Other long-term care (4)	852	0
College dormitory (5)	208	5
Military facility (6)	136	1
Other non-institutional (7)	1,530	2

Table 11: True and Expected Number of GQs by Year

	County A		County B		County C	
Year	True	Expected	True	Expected	True	Expected
0	2	–	6	–	19	–
1	2	2	6	6	15	19
2	1	2	4	6	15	19
3	0	2	4	6	12	18
4	0	2	4	6	11	17
5	0	2	4	6	11	17

4.5 Assumptions About the Expected Counts

Among the GQs which are not in sample there are two important, but unknown, attributes: (1) whether or not the GQ exists, and (2) how many people live in the GQ. Both are generally updated only for GQs *in sample*. Nevertheless, we rely heavily on both facts in the imputation and weighting steps. For example, the imputation-based estimates in cases such as those described in Section 4.4 could be further improved given this information. To illustrate the potential effects of each element, we examine three counties in the state of Ohio which contain college dormitories (major type 5).

In Table 11, the true number of GQs in each county for each year is shown along with the expected number. Note that the year 0 row represents the Census 2000 value. In each of these cases, the actual number of GQs differs from the expected number. This is simply because there is no formal mechanism for updating GQs closings if a GQ is not in sample. For County A, we never find out that in years 3-5, there are no college dormitories. Since the GQs in the county are not sampled, their population count is not updated and we impute records into a county which should have a count of 0 for both the three and the 5-year estimates. This is an example of attribute (1).

Element (2) can be partially examined in Table 12. In Table 12 we list, for each year, the true total population of the county along with the expected total population. There are clearly large discrepancies between each pair of columns. This occurs because not only do some GQs close, existing GQs can change size each year and sometimes the changes are extreme as well. Again, as with attribute (1), the counts of a GQ are not updated on the sampling frame if it is not sampled. The tract and county constraints depend heavily on these expected counts and not having current counts available may have a negative impact on the imputation-based estimates.

4.6 Comparing Methods

There is little difference in the results between the "county" methods and the "tract" methods. That is, "expand county" and "expand tract" are similar and "k-means county" and "k-means tract" are similar. These two pairs of methods differ only by which GQs are chosen for imputation, not where

Table 12: True and Expected Population Counts

	County A		County B		County C	
Year	True	Expected	True	Expected	True	Expected
1	38	38	834	843	1,527	684
2	38	38	656	808	1,518	825
3	0	38	713	863	1,796	726
4	0	38	687	931	687	931
5	0	38	700	938	1,502	596

the donors are found. The number of imputations varies little between the "county" and "tract" methods which indicates that most of the difference across methods results from the choice of donor.

The expanding search methods and K-means methods find donors in very different locations. If physical proximity is more important, then the expanding search methods should perform better than the K-means methods. If the characteristics of the surrounding housing population are of greater value, then the converse should be true.

We find that in general, the K-means methods perform no better, and often worse, when compared to the expanding search methods. This is the case when comparing bias and, to a lesser extent, for MSE, MAD, and % MAD. In some cases, however, the K-means methods perform better in terms of lower variance.

One example is shown in Figure 8 for estimates of Hispanic persons (counties, 5-year estimates). The top plot graphs the bias across counties for each GQ type. Each individual box plot represents a method (design-based included). The bottom plot shows standard deviation in a similar way. The K-means methods tend to have a much higher bias compared to the expanding search methods especially for GQ types 1 (correctional facilities), 5 (college dormitories), and 6 (military facilities). However, the variance tends to be a bit lower for the K-means methods.

5 Discussion

While the ACS GQ estimates are published mainly at the state and national levels, GQ data are also combined with the household data to produce tract and county level estimates for the total resident population. Researchers have shown that estimates for various characteristics of the total resident population can change drastically from year to year for small areas. This variation may not be indicative of real changes, but rather an artifact of particular GQs being in sample in one year and not the next. In order to produce more consistent estimates, we propose an imputation-based approach to, in a sense, "fill in" gaps in GQ representation. Whole person records are imputed into select GQs to obtain representation for all major GQ types at the tract level for 5-year estimates, and at the county level for all period estimates. These imputed records are appended to the sampled records to create the final augmented data set.

In this technical report, we propose four imputation-based methods which achieve the following objective: complete county by type group representation for 1- and 3-year estimates and complete tract by type group representation for 5-year estimates. These imputation procedures are constructed by combining two methods of choosing which GQ to impute into with two methods of selecting donors. We choose donors based on physical proximity (expanding search) or by the similarity of the surrounding housing population (K-means clustering). In order to test the methods, we simulate a GQ population using Census 2000 short-form data and draw 25 independent ACS samples from this population. The advantage of this approach is that the true, population value of each variable is available so we can directly compare methods.

The value of testing imputation methods using the simulation study lies almost entirely in the

GQ population construction step. The more realistic the simulated population, the more valuable the conclusions made about the efficacy of the proposed methodology are. In the simulation design process, a concerted effort is made to make it as realistic as possible. Specifically, we allow for the GQ population to evolve over the five years by implementing three key features:

1. The number of residents in a GQ can change across years.
2. GQs are allowed to close.
3. College dormitories in the summer are set to have no residents. This element is to introduce a notion of seasonality inherent in the GQ population into the simulation.

These features, however, are rough approximations of what happens in reality. Moreover, we do not allow for any changes in the composition of GQ resident characteristics and do not add any new GQs to the universe. These are all critical limitations of the simulation study. The results described in this paper, however, are still useful both in assessing the general performance and in identifying problems in the proposed imputation methodology.

We find, for instance, nearly half of the augmented data are comprised of imputed records, regardless of the imputation method applied. In addition, the number of imputed records can far exceed the number of sampled records for some major GQ types. Most donors are found within the specific GQ type across all imputation methods. While the K-means methods generally find donors outside of the state, the expanding search methods generally find donors within the state, and many times within the county of the GQ to be imputed.

We also discover that the imputation-based methods are systematically biased even at the state level, at times considerably so. The variances are smaller than the design-based estimate variances and comparisons of the MSE, MAD, and %MAD give mixed results. The two ways of choosing GQs to impute into yield similar results; however, the K-means methods seem to perform less well than the expanding search methods. Nevertheless, some of the downsides of the imputation-based methods can be mitigated by changes to the imputation methodology.

There is one critical shortcoming to the proposed imputation methods: they rely heavily on information about GQs which are not in sample. Each piece of information about an unobserved (not in sample) GQ such as the number of residents, the GQ type, the composition of its residents, or whether it even exists may be inaccurate on the sampling frame. For instance, we may impute persons into nonexistent facilities as a result of these sampling frame issues. Therefore, any major improvements must be targeted at learning more about each individual GQ.

Better information about a GQ can lead to a better donor selection procedure. The goal here is to find the best possible donors for each GQ requiring imputation. Four areas where tangible improvement in the imputation procedure can be made are identified next. These changes will be implemented in the second phase of the larger ACS GQ collaborative project.

The major and specific GQ type designations do not indicate whether a GQ is single-sex or mixed. However, the simulation study shows that estimates for the number of female residents and the number of male residents are biased especially in correctional facilities. The sex ratio of individual GQs could be computed from auxiliary information, if available, and used to select donor pools which reflect the sex ratio in the imputed GQ. For instance, a GQ thought to be all female, would have only female donors imputed into it, not male donors. Such a procedure would help reduce the bias of the estimates. Fortunately, we can compile this type of information from sources such as historical ACS sample records and census records.

A second problem with the current methodology is that the expected number of residents in a GQ is not a good approximation of the true, but unobserved population size. These expected values do not account for the changes in population size over time. Furthermore, the delete rates constructed for the simulation are rough approxmations based on observed ACS data and account only for GQ closings. In reality, however, GQs both close and open. As there is no systematic procedure in place to incorporate GQs openings, the delete rate which is applied to the simulation is artificially high. However, both the imputation and weighting steps depend heavily on the expected population sizes and delete rates. Therefore the proposed methodology could be improved by constructing an algorithm which estimates these quantities more accurately.

ACS estimates are designed to incorporate seasonal changes in local populations. For instance, college dormitories are sampled throughout the year even though there may be fewer students living there during the summer. In the current imputation procedure, college dormitories which are sampled in the summer are designed to have no residents and as a result could be candidates for imputation. In particular, if a GQ is large and no residents are sampled, they will be automatically imputed. This imputation step disregards the seasonal aspect of the college dormitory population. If we restrict imputation of such GQs to only the cases where the imputation is required to achieve tract or county by major GQ type representation, we can avoid this issue.

A final, more technical, issue is that of repeating donors. At present, the number of times a donor can be used is limited *overall*. Applying this rule still, however, allows for the possibility that a donor is used multiple times within a *tract*. In such a situation, the donor's characteristics are highly concentrated in one area. This is undesirable especially if the chosen donor is atypical. Therefore, it may be better to limit donor repetition within a tract while still maintaining an overall, more global, maximum. Note that this bound must be chosen with care. The desire for diluting individual donor influence may go against the desire to find donors who are geographically close.

Each of the seven major GQ types house people who have very different characteristics from each other and from household residents. Therefore, especially for small areas, obtaining information on GQ residents is vital to producing estimates of the total resident population. The ACS GQ sample and weighting design, while adequate for producing state-level estimates, is unsatisfactory at sub-state levels of geography. Counties and tracts with no GQ samples despite containing GQs are treated as having no GQ residents at all. The imputation-based methods discussed here rectify this problem by imputing GQ residents into these unrepresented areas.

Acknowledgments

We would like to thank Mark Asiala, Michael Beaghen, Patrick Joyce, Alfredo Navarro, Yves Thibaudeau, and Lynn Weidman for their contributions to this project and their helpful suggestions on this report.

References

Mark Asiala. *Proposal for 2006-2010 5-year GQ Estimation.* Workshop on ACS Group Quarters Small Area Estimation, Mar. 17, 2010.

Nancy Bates and Mary H. Mulry. Segmenting the population for the census 2010 integrated communications program. *C2PO 2010 Census Integrated Communications Research Memoranda Series*, pages 1–28, 2008.

Michael Beaghen and Sharon Stern. *Usability of the American Community Survey Estimates of the Group Quarters Population for Substate Geographies.* Proceedings of the 2009 Joint Statistical Meetings on CD-ROM, American Statistical Association: 2213-2137, 2009.

Antonio Bruce and J. Gregory Robinson. *Tract Level Planning Database with Census 2000 Data.* U.S. Government Printing Office, Washington, DC, 2007.

Bureau of the Census. *Design and Methodology: American Community Survey.* U.S. Government Printing Office, Washington, DC, 2009.

John A. Hartigan. *Clustering Algorithms.* Wiley, New York, 1975.

Patrick M. Joyce. *Constructing Synthetic Group Quarter Populations for Facilitating the Evaluation of Imputation-Based Estimation Methods for the American Community Survey: (Draft Report)*. 2010.

Chaitra H. Nagaraja. *01 Evaluation Summary (Draft Report)*. 2010a.

Chaitra H. Nagaraja. *02 1-year Evaluation Summary (Draft Report)*. 2010b.

Chaitra H. Nagaraja. *03 3-year Evaluation Summary (Draft Report)*. 2010c.

Chaitra H. Nagaraja. *04 5-year Evaluation Summary (Draft Report)*. 2010d.

Chaitra H. Nagaraja. *06 Evaluations Part II (Draft Report)*. 2010e.

Andre L. Williams. *Specifications for selecting the American Community Survey group quarters sample*. #ACS09-S-6, 2008.

A Tables

Table 13: Region, Division and Federal Information Processing System (FIPS) State Codes

Northeast Region (1)	
New England Division (1)	**Middle Atlantic Division** (2)
Connecticut (9)	New Jersey (34)
Maine (23)	New York (36)
Massachusetts (25)	Pennsylvania (42)
New Hampshire (33)	
Rhode Island (44)	
Vermont (50)	
Midwest Region (2)	
East North Central Division (3)	**West North Central Division** (4)
Illinois (17)	Iowa (19)
Indiana (18)	Kansas (20)
Michigan (26)	Minnesota (27)
Ohio (39)	Missouri (29)
Wisconsin (55)	Nebraska (31)
	North Dakota (38)
	South Dakota (46)
South Region (3)	
South Atlantic Division (5)	**East South Central Division** (6)
Delaware (10)	Alabama (1)
District of Columbia (11)	Kentucky (21)
Florida (12)	Mississippi (28)
Georgia (13)	Tennessee (47)
Maryland (24)	**West South Central Division** (7)
North Carolina (37)	Arkansas (5)
South Carolina (45)	Louisiana (22)
Virginia (51)	Oklahoma (40)
West Virginia (54)	Texas (48)
West Region (4)	
Mountain Division (8)	**Pacific Division** (9)
Arizona (4)	Alaska (2)
Colorado (8)	California (6)
Idaho (16)	Hawaii (15)
Montana (30)	Oregon (41)
Nevada (32)	Washington (53)
New Mexico (35)	
Utah (49)	
Wyoming (56)	

Table 14: Number of Tracts in Each Cluster by State

	Cluster							
State	1	2	3	4	5	6	7	8
Alabama	544	77	217	27	12	0	27	159
Alaska	33	47	3	0	3	0	16	24
Arizona	256	214	30	15	109	40	62	295
Arkansas	378	50	114	5	9	0	6	51
California	1,410	1,130	232	109	985	735	566	1,683
Colorado	303	192	34	8	37	8	79	349
Connecticut	175	157	1	84	0	11	54	319
Delaware	85	30	12	6	1	0	3	47
District of Columbia	15	28	34	50	0	2	39	7
Florida	1,062	582	243	81	122	76	156	673
Georgia	643	184	270	73	18	9	80	307
Hawaii	74	49	6	3	9	12	33	74
Idaho	145	32	7	0	3	0	8	67
Illinois	808	331	164	263	55	126	251	903
Indiana	505	143	165	34	5	1	55	477
Iowa	314	74	21	11	1	0	18	344
Kansas	268	82	63	6	14	1	25	257
Kentucky	469	99	119	25	76	1	22	165
Louisiana	448	112	283	42	21	0	33	134
Maine	150	49	1	6	1	0	19	58
Maryland	277	183	138	27	4	3	80	470
Massachusetts	289	268	6	121	1	32	214	371
Michigan	754	271	335	64	18	3	95	975
Minnesota	388	176	31	28	0	3	63	543
Mississippi	296	38	188	7	1	0	8	56
Missouri	573	125	160	48	9	0	63	295
Montana	156	27	19	1	0	0	9	28
Nebraska	191	54	26	4	6	1	22	188
Nevada	129	102	10	8	8	18	58	130
New Hampshire	77	46	0	4	0	0	24	94
New Jersey	371	294	70	134	3	107	167	725
New Mexico	162	76	43	4	51	1	12	65
New York	1,056	794	86	705	19	396	610	948
North Carolina	785	179	216	46	4	0	59	200
North Dakota	104	34	9	1	0	0	7	54
Ohio	945	367	347	152	10	1	123	929
Oklahoma	480	93	156	11	10	0	36	179
Oregon	338	165	30	4	6	2	49	136
Pennsylvania	1,126	381	361	75	15	1	93	963
Rhode Island	56	52	0	36	1	5	15	58
South Carolina	482	68	141	19	1	0	26	84
South Dakota	94	31	31	1	0	0	8	50
Tennessee	598	114	188	50	18	0	48	217
Texas	1,393	512	337	55	718	128	311	788
Utah	110	65	11	6	7	0	29	242
Vermont	87	25	0	0	0	0	9	32
Virginia	511	216	139	33	8	7	91	484
Washington	431	272	45	13	26	9	111	376
West Virginia	314	17	49	11	15	0	5	43
Wisconsin	446	224	36	58	0	15	74	372
Wyoming	70	26	3	0	0	0	2	18
Total	21,174	8,957	5,230	2,574	2,440	1,754	4,073	16,506

B Figures

Figure 2: Counting Records: Military Facilities (6)

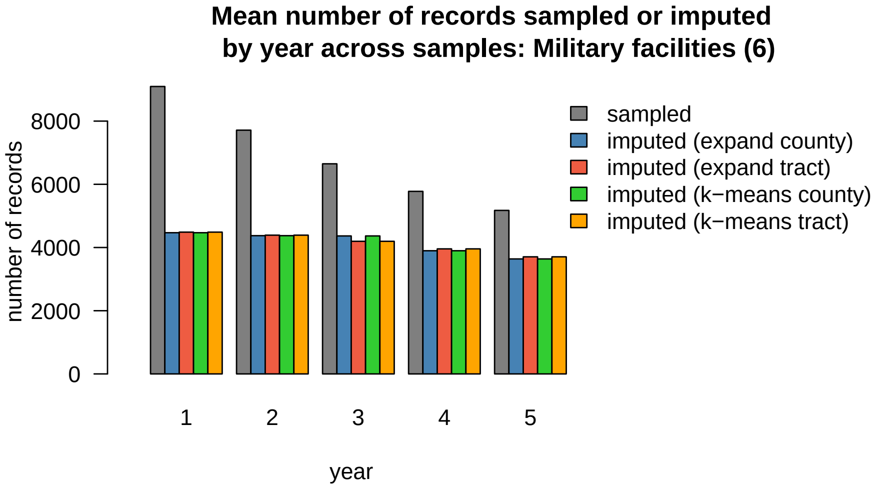


Figure 3: Counting Records: Juvenile Detention Facilities (2)

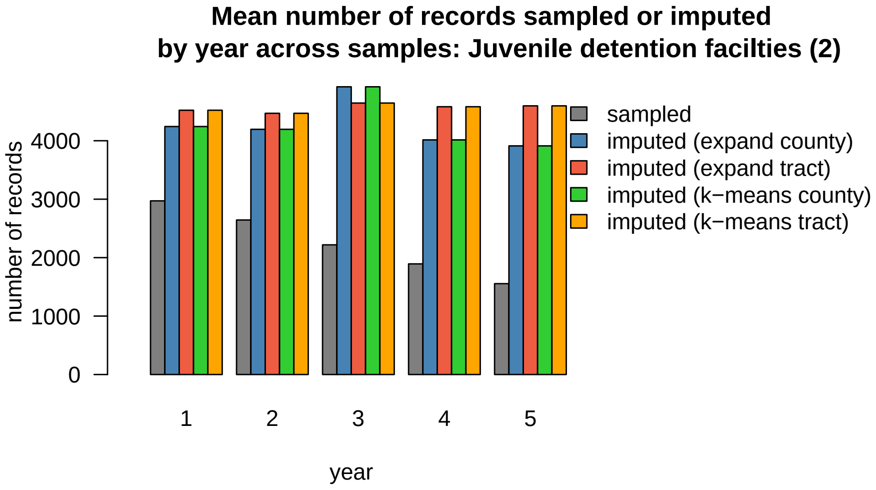
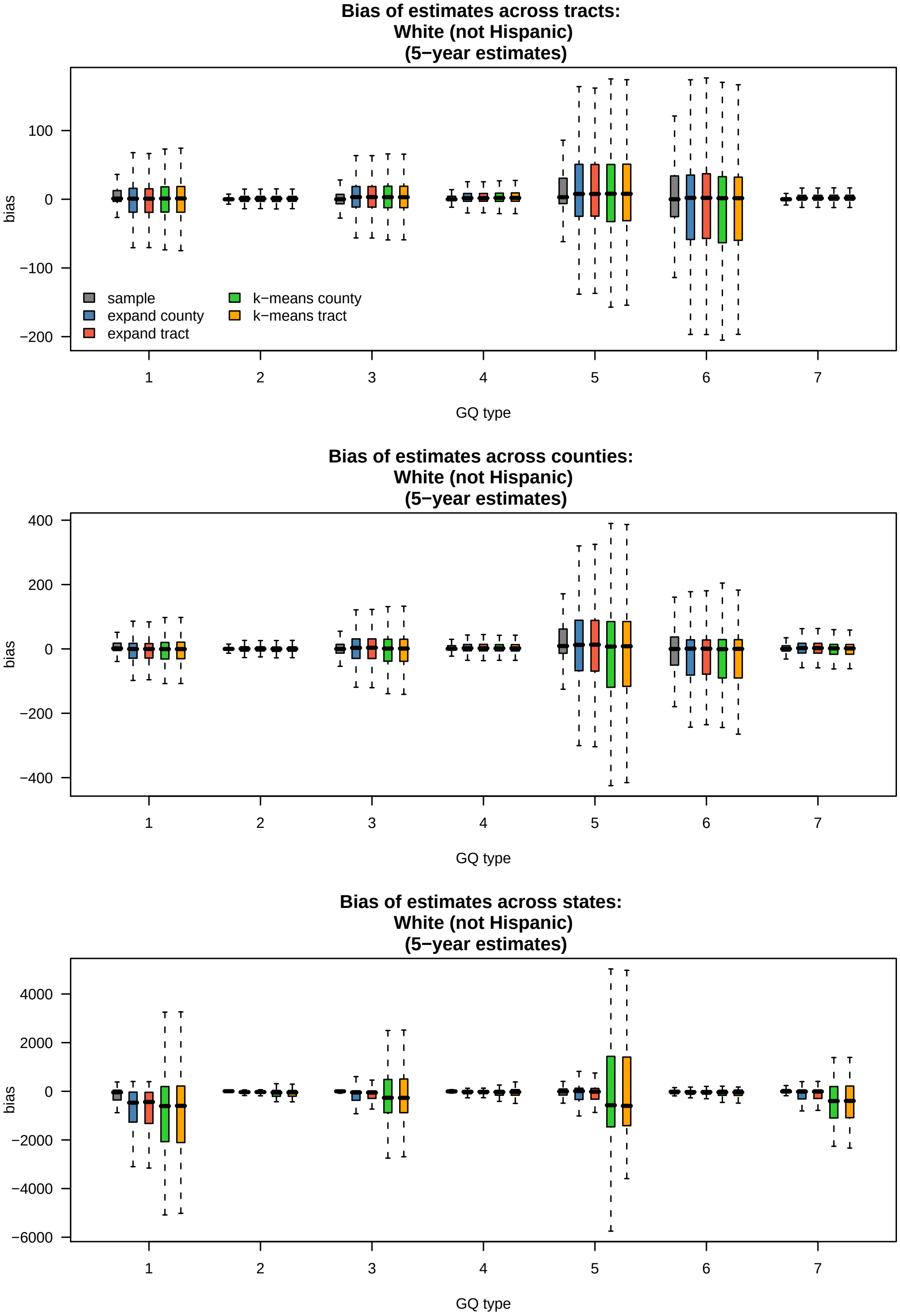


Figure 4: Bias at the Tract, County and State Levels (5-year estimates)



Bias of estimates across tracts: White (not Hispanic) (5−year estimates)

Bias of estimates across counties: White (not Hispanic) (5−year estimates)

Bias of estimates across states: White (not Hispanic) (5−year estimates)

Figure 5: Design-based Estimates Against Imputation-based Estimates

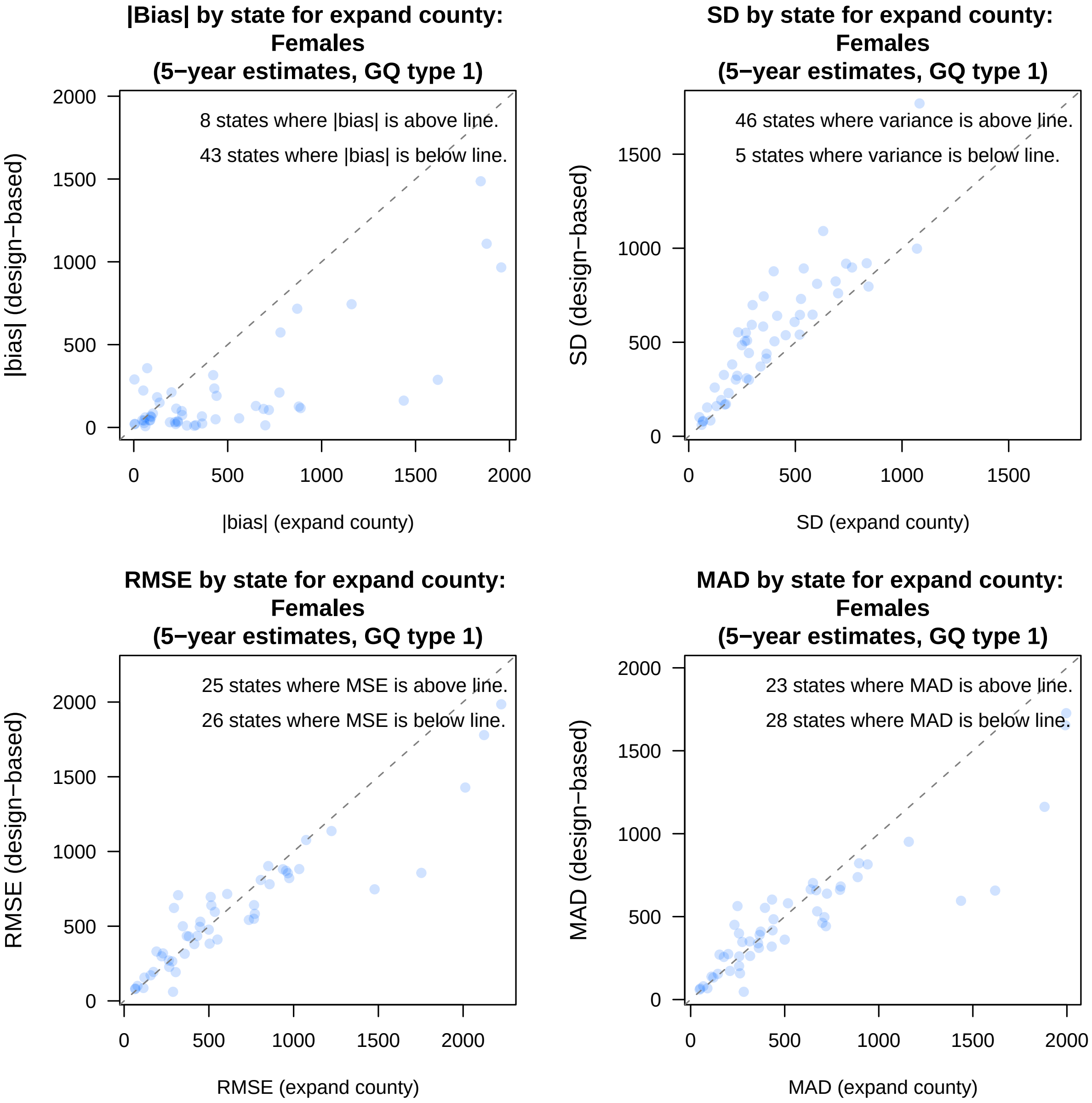


Figure 6: Scatterplots of Summary Measures

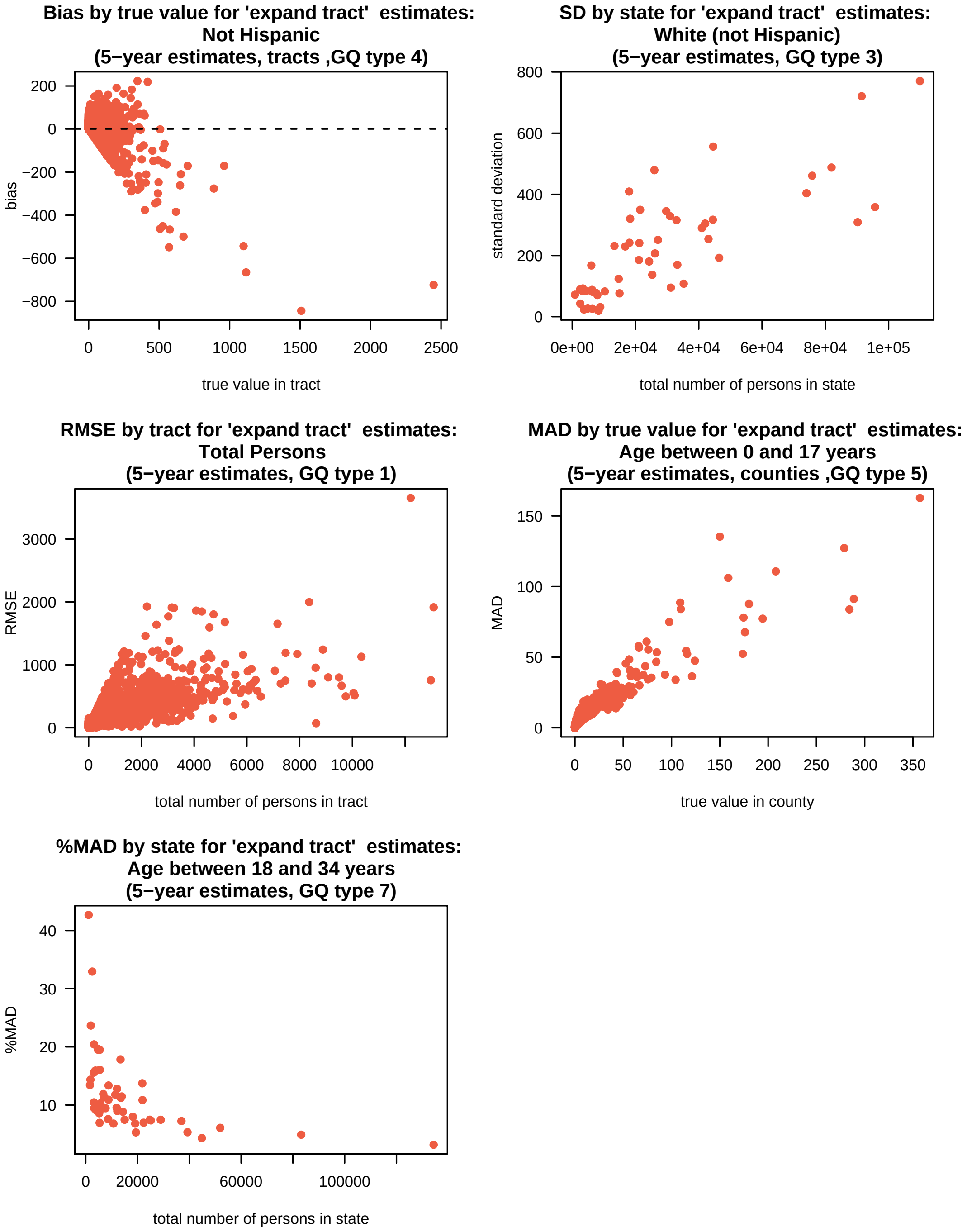


Figure 7: Comparing Geographies with No Sampled or No Imputed Records

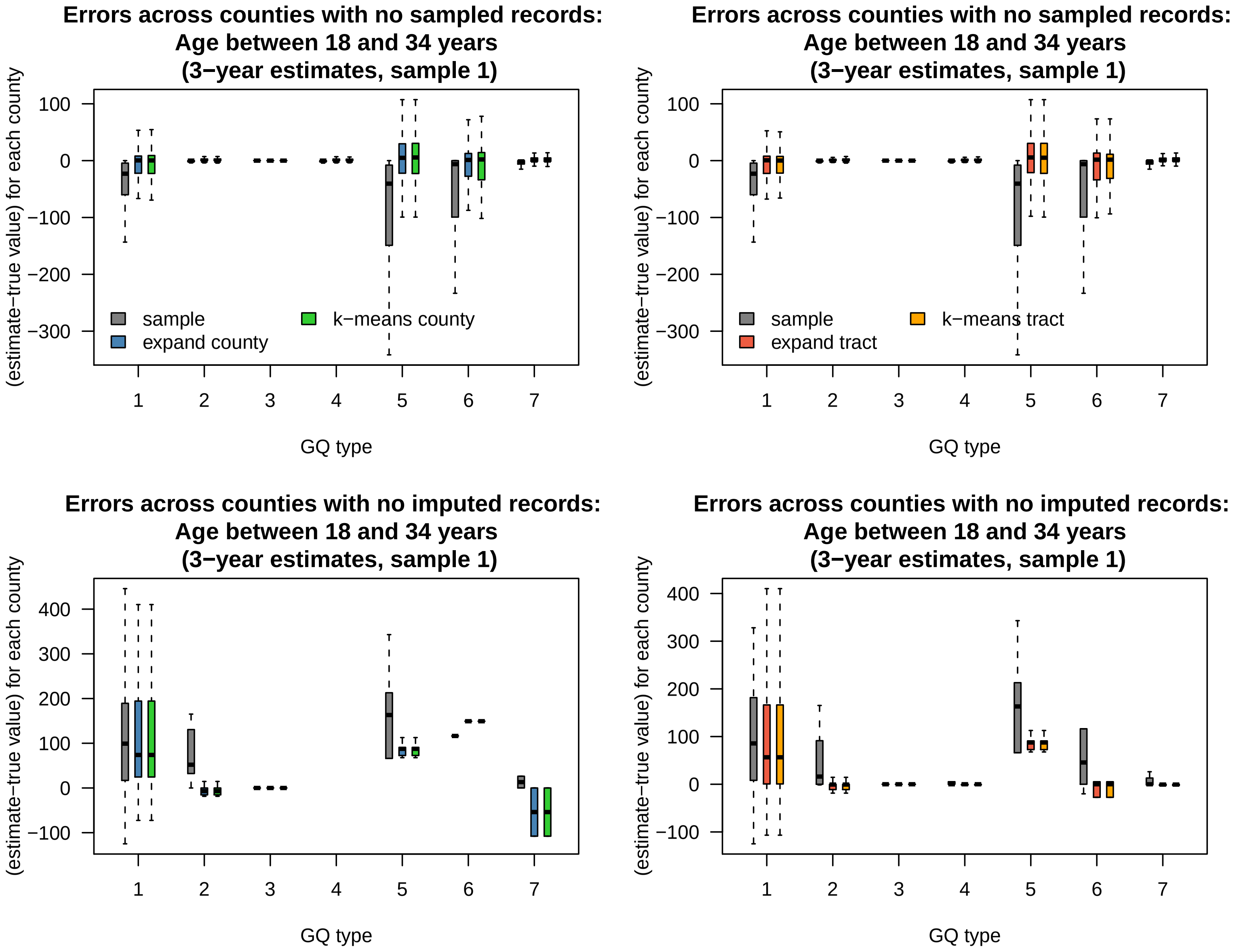
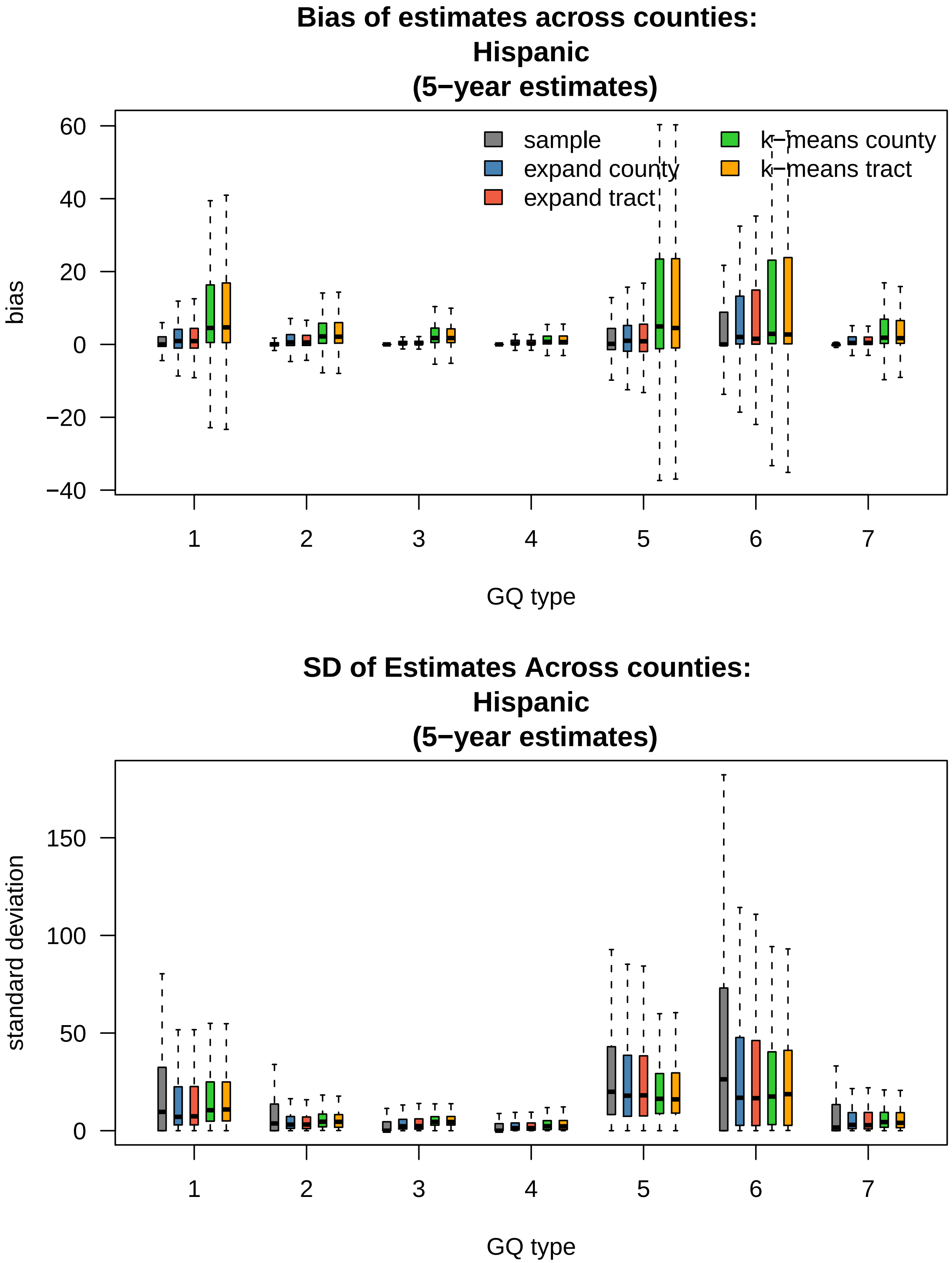


Figure 8: K-means Versus Expanding Search Donor Selection Methods


Bias of estimates across counties:
Hispanic
(5−year estimates)
bias
sample
expand county
expand tract
k−means county
k−means tract
GQ type
SD of Estimates Across counties:
Hispanic
(5−year estimates)
standard deviation
GQ type