10/08/2020

Beatrix Haddock

Simulation Science

# Maintaining accuracy while protecting privacy in the 2020 Decennial Census: Binning as a strategy to reduce the homogeneity index

## Introduction

The Census bureau operates under a constitutional mandate to enumerate every person living in the US every ten years. This decennial count is used not only to allot congressional representation amongst the states, but for apportioning funding at the federal and state level. It is also used extensively in social surveying and other social research. As a data source, the US decennial census is unparalleled in its breadth and level of detail, and has become indispensable to {list examples: urban planning, housing development, voting rights, and the sociological investigation of education, crime, family, and race, among other things}. However, as stewards of such a large database of sensitive information, the Census Bureau is additionally mandated by Title 13 of US Code to not publish any information that allows the re-identification of any specific individual. In order to comply with this privacy mandate, this year the Census is overhauling their privacy protection methods that have become insufficient in lieu of a Differentially Private program.

The degree of privacy offered by any disclosure avoidance system is inversely related to the quantity and accuracy of information the system reveals about the individuals described in the database. The Census Bureau has been working {since X} to develop a program that balances privacy and accuracy; however, as the first public sector data provider to implement Differential Privacy, the Bureau has encountered significant challenges both in maintaining the accuracy needed by its data user community, and mitigating the bias that their program introduces.

One challenge in producing a differentially private version of the decennial census is the massive number of “histogram bins” in the fully saturated contingency table of counts of the US population stratified by location, age, sex, race, ethnicity, and relationship to head-of-household. [[brief summary of Georgetown effort to reduce number of bins.]] The hypothesis behind this binning is that it will be introduce less error to guarantee privacy for less bins. Although this can only be tested definitively with access to the non-private census data currently being gathered, Census Bureau has agreed to also investigate it with their non-private 2010 decennial census data. It is also possible to use publicly available data and population synthesis to reconstruct a population similar to that we expect to respond to the 2020 census and to examine binning data as a strategy to reduce the quantify of noise needed for a given level of privacy with this reconstructed population. In this paper, we develop one such population synthesis approach to predicting impact of binning strategies, and provide estimates of the potential improvement in error that a binned disclosure avoidance system might offer.

## Methods

In this section, we provide (1) a concise summary of differential privacy, the TopDown algorithm to guarantee privacy in 2020 decennial census; (2) a description of the binning strategies developed by [[Georgetown group]] to potentially reduce the error introduced by TopDown; (3) our reconstruction of the 2010 Decennial Census person data; (4) our approach for predicting impact of binning on homogeneity index.

Differential privacy and the TopDown algorithm

**Definition: Differentially Private.** Let A be an algorithm, and D\_1 and D\_2 databases that are identical everywhere except for one entry. Then A is epsilon-differentially private if:

Pr(A(D\_1)=P) <= exp(epsilon) \* Pr(A(D\_2)=P)

Intuitively, we can see that from looking at data product P, if we cannot tell if it came from a database that looks like D\_1, D\_2, or any other D\_n which differs from D\_1 on one person, then we cannot draw conclusions about any single individual from the information in P.

The Census bureau divides the US population into a nested, 7-level hierarchy: National, State, County, Tract, Blockgroup, Block. The bulk of the data products Census will release will be in the form of tabulated counts of individuals of a certain race, ethnicity, sex, age, household type, and household relationship for a given geographic unit in their geographic hierarchy. Census Bureau has developed a differentially private algorithm, called TopDown, to produce consistent tabulated counts within a specified privacy loss budget, epsilon. TopDown is still a work-in-progress, but, as currently implemented, it iteratively adds {Laplace or **geometric**?} noise to counts at successive descending levels of the hierarchy, followed by optimizing the counts at each level to be internally consistent with the higher levels and as close to the noisy counts as possible. This generates final, synthetic counts that meet certain criteria, including:

1. All counts are nonnegative integers;
2. The number of housing units in each block is precise;
3. The number of occupied group quarters is precise;
4. The population of each state is precise.
5. Certain logical requirements regarding the age gaps between parents and children, the counts and ages of householders, and the sexes and ages of certain group quarters units.
6. This was from the E2E publication; need to determine if they’ve made any further commitments or suggestions for their final set of invariants and structural zeros

In their paper [title], Petti and Flaxman found that the bias introduced via this optimization is positively associated with the homogeneity index of the underlying data. They defined the **detailed histogram** to be all the precise counts at the most detailed level of stratification across all variables of interest, and defined the **homogeneity index** of the data to be the number of counts in the underlying detailed histogram that equal exactly zero. [P/A paper] found the bias introduced by TopDown to be positively associated with the homogeneity index (since that paper examined a DP version of the 1940 decennial census, the bias was calculated for each “enumeration district”, a geographic area similar in size to a blockgroup in recent decennial censuses; it found that [[summary of size of relationship between homogeneity index and bias as a function of epsilon]].

In this paper, we will consider a generalization of the homogeneity index of Petti and Flaxman, which we term the n-homogeneity index (n-HI); for a given detailed histogram, we define the n-HI to be the number of counts less than or equal to n. Note that the HI from Petti and Flaxman’s work is thus the 0-HI in our generalized notation.

We find this generalization useful for two reasons. First, as TopDown has the potential to push small non-zero counts negative, (we anticipate) n-HI for an appropriately chosen value of n to more accurately predict bias introduced by TopDown. Second, n-HI captures information about the effects of adjusting epsilon, by taking as an input the distribution of small-value counts in the detailed histogram.

This extended measure allows us to calculate the anticipated degree of bias for a given selection of parameters and adjustments to Census’s disclosure avoidance system.

## Binning strategies that might reduce the error introduced by TopDown

By binning, we mean aggregating groups of variables and their respective counts. For example, we could take the following single-year age counts: {age 90: 4, age 91: 0, age 92: 0, age 93: 1, age 94: 0} and bin them as follows: {age 90-94: 5}. Observe that binning will never increase n-HI, and very likely reduce it. However, binning also loses detail of information.

We must proceed assuming that any binning in Census’s DAS is irreversible. This follows from Census’s need to balance privacy and accuracy of information, and the fact that a future release of unbinned data from the 2020 Decennial would degrade the level of privacy in the collection of all 2020 Decennial Census data products. [Example: table with cross tabs of 7-race and 5yr-age; let there be only one] It is thus of utmost importance to engage all {external stakeholders/the full data user community} to ensure that no such party relies on the information {we} propose to bin.

The following age-race binning scenarios are the proposals settled on by {attendees of a virtual convening (roundtable?) held by MALDEF, AAJC, and Georgetown}, targeted at determining what detail of information their work could afford to relinquish. We additionally include a relationship-binning scenario suggested in consultation with {Georgetown/MALDEF?}:

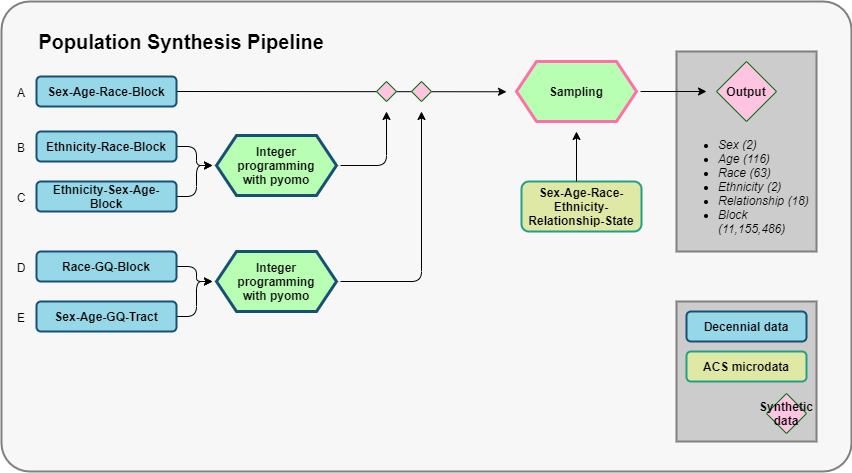
* ­­Age Binning Scenarios
  + A1: {1, 2, … 83, 84} + {85-89, 90-94, 95-99, 100+}
  + A2: {1, 2, … 88, 89} + {90-94, 95-99, 100+}
  + A3: {1, 2, … 88, 89} + {90+}
* Race Binning Scenarios
  + RC1: bin multiracial individuals of four or more races
  + RC2: bin multiracial individuals of three of more races
* Relationship binning scenario:
  + {Head of household, relative of same generation of head of household, relative of parent generation of head of household, relative of child generation of head of household, relative of grandchild generation of head of household, nonrelative}

Population synthesis and the impact of binning on homogeneity index

One could obtain the most accurate assessment of these strategies with access to the true, unprotected census data, and a fully up-to-date version of the TopDown program. As such, [MALDEF, AAJC, Georgetown et al] proposed combinations of the above age and race binning scenarios to the Census Bureau for testing and implementation. However, the Census Bureau is both working under extreme time pressure, and with stringent provisions against publishing additional information synthesized from Census microdata. These factors might prevent Census from flexibly incorporating as much iterative feedback from the data user community into their testing and decision-making process as we would wish for. [Note on why it is important for data user community to be involved in this process.]

## US population synthesis

To understand the comparative effects of our binning scenarios, it is necessary to have data with an accurate joint distribution across the variables under consideration for binning. Here we discuss the synthesis of such a microdata file. [It is of note that our ability to reconstruct such a person file is the driving incentive for the Bureau’s introduction of a differentially private disclosure avoidance system.]



Our synthetic population relies on the following Census data products:

1. From the 2010 Decennial Summary File 1 (SF 1):
   * **Decennial A:** block-level counts of individuals cross-tabulated by:
     + Sex, 2 values
     + Race, 7 values
     + Age, 23 values
   * **Decennial B:** block-level counts of Hispanic individuals, cross-tabulated by:
     + Race, 7 values
   * **Decennial C:** block-level counts of Hispanic individuals, cross-tabulated by:
     + Sex, 2 values
     + Age, 23 values
   * **Decennial D:** block-level counts of individuals cross-tabulated by:
     + Race, 6 values
     + Group quarters status (institutionalized GQ, non-institutionalized GQ, non-GQ)
   * **Decennial E**: Tract-level counts of individuals, cross-tabulated by:
     + Sex, 2 values
     + Age, 3 values
     + Group quarters status (institutionalized GQ, non-institutionalized GQ, non-GQ)
2. **ACS:** From the 2012-2017 1-year ACS surveys, State-level counts, cross-tabulated by:
   * Sex, 2 values
   * Age, 116 values
   * Race, 63 values
   * Ethnicity, 2 values
   * Relationship to head of household, 18 values

Our 2-sex, 6-race, 23-age, block-level tabulations serve as our baseline framework. Separately, we use our data on the Hispanic population to synthesize cross-tabulations meeting the following constraints:

1. The 7-race-stratified counts are accurate, given by Decennial B
2. The 2-sex, 23-age counts are accurate, given by Decennial C
3. The 2-sex, 23-age, 7-race counts of Hispanic individuals, per block, are capped by the 2-sex, 23-age, 7-race all-ethnicity counts per block, given by Decennial A.

We do this by initializing empty matrices that represent the space {46 sex-age} x {7-race}, such that each cell represents a potential age-sex-race combination, and optimizing using Pyomo to synthesize counts meeting our constraints.

We then subtract the Hispanic population out from the all-ethnicity population to obtain non-Hispanic counts, thereby creating synthetic block-level counts of individuals cross-tabulated by:

1. Sex, 2 values
2. Age, 23, values
3. Race, 7 values
4. Ethnicity, 2 values

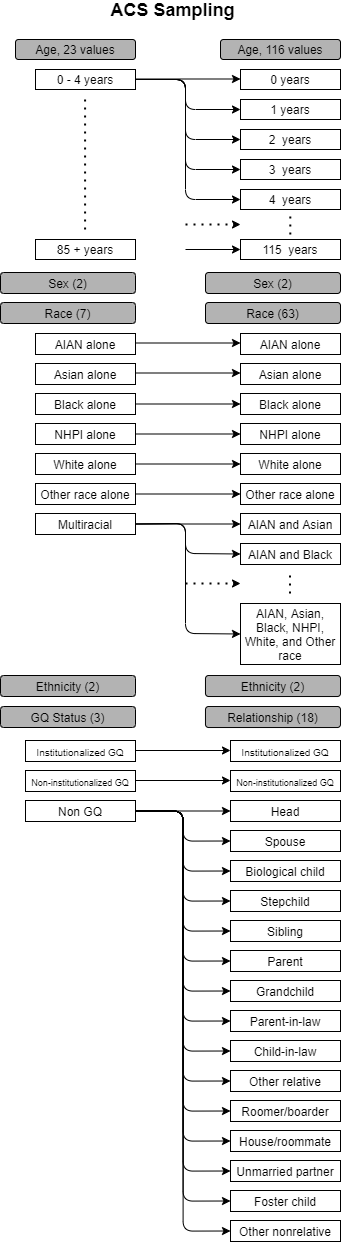
We next stratify our counts by group quarters status. For each tract, we initialize an empty matrix representing the space {3-group quarter status} x {6 sex-age} x {7-race} x {n-block}, with n equaling the number of blocks in our tract. Using Pyomo, we populate this matrix to meet the following constraints:

1. The 7-race counts stratified by 3-GQ status are accurate at the block level, given by Decennial D
2. The 2-sex, 3-age counts stratified by 3-GQ status are accurate at the tract level, given by Decennial E
3. The sum of individuals in institutionalized GQ and non-institutionalized GQ, for each 2-sex, 3-age, 7-race category, is bound by the total number of individuals in each 2-sex, 3-age, 7-race category at the block level, given by Decennial A.

From this process we obtain a table for each block, with counts of individuals in group quarters stratified by:

1. GQ status, 2 values
2. Sex, 2 values
3. Age, 3 values
4. Race, 7 values

We then define a map from our GQ dataset to our synthetic sex-age-race-ethnicity dataset, which assigns each individual in the GQ dataset a more specific age. To all other individuals in our synthetic dataset, we assign a GQ status of “non-group quarters”, thereby obtaining data stratified by:

1. Sex, 2 values
2. Age, 23, values
3. Race, 7 values
4. Ethnicity, 2 values
5. GQ status, 3 values

Our goal, from this stage, is to disaggregate along the following variables:

1. Age, 116, values
2. Race, 63 values
3. Relationship to head of household x GQ status, 18 values

To get this granularity, we use the data from the 1-years ACS surveys from 2012-2017. This is person-level microdata at the state level, from which we keep the following variables:

1. Age, 116 variables
2. Sex, 2 variables
3. Race, 63 variables
4. Ethnicity, 2 variables
5. Relationship to head of household x GQ status, 18 values

For each bin in our synthetic sex-age-race-ethnicity-gq dataset with count ni > 0, we subset our ACS data to all individuals in the corresponding State who match on sex, age, race, ethnicity, and GQ status. If no such individuals exist, we start dropping matching requirements beginning with sex, then ethnicity, then race, then GQ status.

From this pool, we sample ni specific individuals with uniform probability and replacement; the individuals in our synthetic bin then inherit the specific age, sex, race, and relationship to head of household characteristics from the ACS individuals. Note that this approach guarantees that our simulants have relationships to head of household that are logically consistent with their own age, sex, race, ethnicity, and GQ status. However, by sampling our relationships to head of household from State-level data, we lose any priors that are independent of age, sex, race, and ethnicity for within-state geographic variation in the distribution of relationships. As we use this method to assign specific racial identities to multiracial individuals, we have this same issue for the distribution of the multiracial population. Observe that could conceivably generate a block in which there is an AIAN and white individual with status “biological child”, and no potential AIAN parents in the same block.

From this process we obtain our final block-level outputs, stratified across the following variables:

1. Sex, 2 values
2. Age, 116, values
3. Race, 63 values
4. Ethnicity, 2 values
5. Relationship to head of household x GQ status, 18 values

## Limitations

By sampling our relationships to head of household from State-level data, we lose any priors that are independent of age, sex, race, and ethnicity for within-state geographic variation in the distribution of relationships. Furthermore, beyond the state level, we use no geographic priors for the distribution of specific types of multiracial individuals.

{Add more limitations}

## Extensions

In order to improve the within-state geographic distribution multiracial individuals, ideally the information on “{Race x} alone and in combination”, available in the 2010 Decennial SF1, should be utilized. Similarly, the multitude of variables informing the relationships to head of household extant in each block would ideally be utilized to generate a more accurate distribution of relationships in each block. Extending this further, we are very interested in grouping individuals into households with appropriate within-household variable distributions, and we believe there is sufficient data available in the 2010 Decennial SF1 to do this. This work could also be improved by updating the ‘relationship to head of household’ variable to contain the variety of information intended for the 2020 Census; however, many of the values in the 2020 iteration would require inference, and we have not investigated how much confidence we could have in the accuracy of a solution for this variable.