Introducing the Decennial Census Total Uncertainty Analysis Project

September 20, 2024

Prepared by
Joe Schafer (U.S. Census Bureau)
for the International Total Survey Error Workshop (ITSEW 2024)

Views expressed are those of the author and not necessarily those of the U.S. Census Bureau. Joseph.L.Schafer@census.gov

United States® Census Bureau

Some Background

- The new, formally private Disclosure Avoidance System (DAS) implemented in the 2020 Census has injected random noise into published Census counts at smaller geographies
- As the Bureau strives to educate stakeholders by quantifying the effects of DAS, we must acknowledge that published Census counts have always been subject to many kinds of non-sampling error
- Estimates of coverage error (erroneous enumerations and omissions) were provided by the 2020 Post-Enumeration Survey (PES), but impacts of other sources of error have not been quantified to the same extent
- Leveraging new data sources and computing capabilities, we can now attempt to model the combined effects of nonsampling error on published Census statistics at varying levels of geography

United States® Census Bureau

Total Uncertainty Analysis (TUA) Project

- **Overall purpose:** To quantify uncertainty in results from the Decennial Census accruing from as many sources as we can within a reasonable time frame. Not only to assess what happened in 2020, but to inform decisions about 2030 and beyond
- **Who is involved:** Joint research effort by staff with expertise in statistical modeling; Census operations, data processing, and imputation; Census evaluation studies and the Post-Enumeration Survey; use of administrative records; research computing. *Louis Avenilla, Andrew Keller, Timothy Kennel, Ryan King, Brian Knop, Rafael Morales, Tom Mule, James Noon, Aneesah Williams, Julianne Zamora*
- **History:** Project began during the 2020 Census and was originally led by Paul Biemer (RTI). After PB's retirement, the work has continued internally at the Bureau
- **Theoretical framework:** Historically, uncertainty was framed in frequentist terms as the **variation over hypothetical repeated censuses** of same population under similar conditions. We have now shifted to a **Bayesian** view, describing posterior uncertainty in unknown true census roster given the observed data.
- **Current state of project:** Creating a data product (first version late 2024) to represent random draws from a posterior distribution of **integer person counts at the block level** within categories of age, sex, race/Hispanic origin, and **counts of housing units (HUs)** by occupancy status and tenure. Product will be for **internal use**, but findings/summaries of uncertainty at varying levels of geography will be made available to public, subject to disclosure avoidance protocols
- **Group quarters:** No plans to describe uncertainty in GQ population, due to limited information on GQ data quality

United States® Census Bureau

Remainder of this Presentation

1. **Sources of uncertainty:** Types of error that contribute to uncertainty in published Census statistics
2. **Inputs:** Sources of data for the TUA project
3. **Modeling:** Constructing a joint model for the “true” and “observed” census; flexible methods and software for fitting multilevel latent-class models with a new R package **bigLC**
4. **Computation:** How we will create the data product using **three stages of multiple imputation**

United States® Census Bureau

1. Types/Sources of Error in Published Census Statistics

1. **Content errors / errors in characteristics:** Discrepancies in recorded or imputed age, sex, race, origin and tenure relative to what the person or a knowledgeable surrogate (e.g., parent) would report under ideal conditions
2. **Errors in Master Address File (MAF):** Discrepancies relative to a complete list of units habitable on Census Day, with no duplicates and every unit in correct block
3. **Errors in person counts:** Arising from inaccurate or incomplete information from self response, field reports, USPS reports, admin sources, status and count imputation
4. **Errors arising in processing and unduplication.** Mistakes in combining information from multiple responses (Primary Selection Algorithm) and special operations for unduplication of units and persons
5. **Disclosure Avoidance:** Noise added by TopDown algorithm

United States® Census Bureau

1. Types/Sources of Error in Published Census Statistics

	Addressed in current activity of TUA?
1. **Content errors / errors in characteristics:** Discrepancies in recorded or imputed age, sex, race, origin and tenure relative to what the person or a knowledgeable surrogate (e.g., parent) would report under ideal conditions	**Partially**
2. **MAF errors:** Discrepancies relative to a complete list of HUs habitable on Census Day, with no erroneous inclusions and every unit place in correct block	**Yes**
3. **Errors in person counts:** Arising from inaccurate or incomplete information from self response, FR reports, USPS reports, admin sources, status and count imputation	**Yes**
4. **Errors arising in processing and unduplication.** Mistakes in combining information from multiple responses (Primary Selection Algorithm) and special operations for unduplication of units and persons	**Yes**
5. **Disclosure Avoidance:** Noise added by TopDown algorithm	**No**

United States® Census Bureau

2. Data Inputs

American Community Survey (ACS)

- Five-year (2016-2020) tract-level estimates of housing and person characteristics which are predictive of 2020 Census data
- Readily available from the 2022 Planning Database, which aligns with 2020 Census geography

2020 Master Address File (MAF)

- Source of 2020 tabulation geography (State, County, Tract, Block) and HU identifiers (MAFIDs)

2020 Census Edited File (CEF)

- Source for official 2020 Census tabulations, prior to noise injected by DAS

2020 Census Unedited File (CUF)

- Census responses from all HUs from the 2020 Decennial MAF, including vacant and delete units, after Primary Selection Algorithm and unduplication
- Also has pre-census variables for HUs (single/multi-unit building, USPS, undeliverable as addressed (UAA), adrec counts…) that were used as predictors for status and count imputation
- HU-level variables (vacant/delete status, self-response, tenure) and final person counts (some imputed)
- Person-level characteristics (pre-edit and imputation)

2020 Post-Enumeration Survey (PES)

Data files that were used to estimate logistic models for 2020 Census coverage

- **E-sample HU model:** predicts erroneous enumerations and mis-locations of housing units
- **E-sample person model:** predicts erroneous enumerations and mis-locations of persons
- **P-sample HU model:** predicts omissions of housing units
- **P-sample person model:** predicts omissions of persons

United States Census Bureau

3. Joint Modeling of True and Observed Census

(Rubin and Zaslavsky, 1989; Zaslavsky, 2004)

"Distributional parameterization"

Factor the joint distribution as

$$P(\text{true census}, \text{observed census}, \theta) = P(\theta) \times P(\text{true census} \mid \theta) \times P(\text{observed census} \mid \text{true census}, \theta)$$

where θ represents all unknown parameters

- Intuitively appealing

"Direct parameterization"

Factor the joint distribution as

$$P(\text{true census}, \text{observed census}, \theta) = P(\theta) \times P(\text{observed census} \mid \theta) \times P(\text{true census} \mid \text{observed census}, \theta)$$

- Appears computationally simpler for TUA
- Aligns more closely with our available data sources
- Aligns more naturally with census coverage estimation as it was implemented in the PES
- Easier to critique models and diagnose lack of fit with respect to the observed census data

United States® Census Bureau

Modeling Census Data

- Our Bayesian framework requires generative probability models describing the joint distribution of Census HU and person counts and characteristics over the landscape
- Previous approaches based on **logistic** and **log-linear models** (Zaslavsky, 2004; Zaslavsky and Zanutto, 2006) don't easily scale up to the large number of observations and variables, nor to the complicated patterns of missing values, appearing in this project
- Census data are **multilevel** (persons within HUs, within blocks, within tracts, …), **multivariate** (multiple intercorrelated variables at each level, with missing values), and **mostly categorical**. Census data also tend to be **clumpy,** with HUs and persons of similar characteristics clustering nearby
- Vermunt *et al.* (2008) and Vidotto *et al.* (2015, 2018a) proposed **latent-class (LC) models with a possibly large number of classes** for multiple imputation of missing values in multivariate categorical datasets
- We call these **predictive LC models** because, unlike traditional applications of LC analysis, the meaning and interpretation of classes are not of main interest
- **Handling large datasets:** Computations for LC models are **linear** in the number of observations, classes, and variables
- With **Markov chain Monte Carlo (MCMC)** and a **Dirichlet Process (stick-breaking) prior**, the effective number of classes can be adaptively chosen by the data (Dunson and Xing, 2009; SI and Reiter, 2013)
- **Nonparametric** in the sense that *any joint distribution* for categorical variables can be represented by an LC model with finite number of classes (Dunson and Xing, 2009)
- **LC models for multilevel data** (Vermunt, 2003, 2008; Hu, Reiter and Wang, 2018; Vidotto, 2018b) can account for clustering and use covariates at higher levels. In a multilevel LC model, each observational unit at each level belongs to a latent class, with class prevalences that vary across the classes at higher levels

United States® Census Bureau

Software for Predictive LC Modeling

Motivation

- Commercial products for LC modeling including **Mplus** (Muthen and Muthen, 2017) and **Latent GOLD** (Vermunt and Magidson, 2016), and existing packages for SAS, Stata and R are not well suited to this project
- We are developing a new R package called **bigLC**
- Designed for predictive LC modeling and imputation of single- and multilevel multivariate data
- Already being used in TUA and several other research projects at the Census Bureau
- After more testing, validation and documentation, we will make **bigLC** available to the public via CRAN and the Census Bureau GitHub

Features

- Designed to handle large datasets; no internal limit to the number of observations, variables, or levels
- Pre- and post-processing coded in R
- Model fitting implemented in Fortran, using the *dotCall:.C64* interface data (Gerber *et al.*, 2018), which can pass long arrays (those having more than $2^{31} - 1 \approx 2.14$ billon elements) and avoids unnecessary copying of data
- Major loops are parallelized using OpenMP, distributing computations over all available CPU cores (up to 8 on a Census Bureau Windows laptop; 64 on a single node of the IRE research cluster) which greatly enhances speed

United States® Census Bureau

4. Three Stages of Imputation

Stage 1: Status, count, and characteristic imputation

Start with the CUF, with tabulation geography added from MAF; wipe out any imputed values produced in the status and count imputation

Multiply impute missing values of

- delete and vacancy status,
- number of persons,
- tenure, and
- person characteristics

under a joint model in a single, integrated procedure.

Result: Multiple versions of a census with coverage errors but no missing values

Stage 2: Imputing erroneous enumeration (EE) status

For each dataset produced in Stage 1, impute EE status using probabilities estimated from PES E-sample regression models

Two versions:

- EE status for HUs
- EE status for persons

Result: Multiple versions of a census of housing units and a census of persons, with no missing values and no erroneous inclusions, but having less-than-complete coverage

Stage 3: Imputing missed HUs and missed persons

For each dataset from Stage 2, impute

- missed HUs in the HU-level file
- missed persons in the person file

using probabilities estimated from the PES P-sample regressions

Uses a new method called **Bayesian expansion,** motivated by the inverse-probability weighting used in the PES

Result: Multiple versions of a census of housing units and a census of persons, with no missing values, no erroneous inclusions, and no omissions

United States® Census Bureau

Stage 1: Status + Count + Characteristic Imputation

Type of model

Version of a **multilevel latent-class model** with up to four levels

Level 1: persons. Items include age, sex, race, Hispanic origin

Level 2: HUs. Items include delete/vacancy status, number of persons, tenure, predictors used for status and count imputation, possibly the characteristics of Person 1

Level 3: Blocks. Number of HUs listed in the MAF

Level 4: Tracts. Number of blocks, estimated vacancy rates, HU density, average number of persons per HU, proportions of persons by race and Hispanic origin from ACS five-year estimates

Partitioning

To keep computations manageable, we are running separate models for counties

Special features of bigLC needed for this model

- Edit constraints to enforce structural zeros (Manrique-Vallier and Reiter, 2014a, 2014b)
- Number of level-1 units (persons), which appears as an item at level 2, is sometimes missing

Separating results into two files

When this integrated imputation procedure is finished, we can separate each imputed dataset into

- **HU-level file** containing HU-level items, with HUs nested within blocks
- **Person-level file** containing person-level variables, with persons nested within blocks. Grouping of persons within HUs won't be necessary for Stage 2 or 3

United States® Census Bureau

Stage 2: Imputing Erroneous Enumeration Status

For each level (HU and person), we will

- Draw parameters of E-sample model from their posterior distribution, then
- Impute an erroneous enumeration status for each HU or person based on the resulting fitted probabilities

Categories

- correct enumeration within block
- correctly enumerated, but mis-located to this block
- duplicate
- other type of erroneous enumeration

Handling mis-locations

For each HU deemed to be mis-located, we will need to impute the correct block location (not difficult; PES says that most of them are from a surrounding block)

For each person deemed to be mis-located, we will need to impute the correct block location (more difficult)

- different block, same county
- different county, same state
- different state

PES created procedures to address this; we can use those procedures as our starting point

United States® Census Bureau

Stage 3: Imputing Missed HUs and Persons

For each level (HU and person), we will

- Draw parameters of P-sample model from their posterior distribution
- Compute omission probabilities for each HU and person
- Use these probabilities to impute omitted HUs and persons, using a new technique of Bayesian expansion

Bayesian expansion

- Given a model for the units that you actually see, and the probabilities of seeing them, draw from the posterior distribution of the unseen units
- Similar in spirit to Horvitz-Thompson expansion (inverse-probability weighted) estimator, adapted to a Bayesian finite-population framework
- Resembles methods for generating synthetic populations from survey data, reversing the complex sample design (Zhou, Elliott and Raghunathan, 2016)
- Like a weighted finite population Bayesian bootstrap (Cohen, 1997; Little and Zheng, 2007; Dong, Elliott and Raghunathan, 2014), but replaces the empirical distribution for the observed units with a parametric model to allow innovation (i.e., may impute an omitted person into a block whose characteristics do not exactly match someone who already exists in the block)

United States® Census Bureau

Bayesian Expansion

- Done within every block
- Done separately for persons and HUs; this is the person version
- Categorical person characteristics (age, sex, race/origin, tenure, …) used in P-sample models for person omissions

$$\boldsymbol{X} = (X_1, \ldots, X_p)$$

- Cells $c = 1, \ldots, C$ of the cross-classified $\boldsymbol{X}$ table
- Total number of persons in cell c is

$$N_c = n_c + m_c$$

n_c = observed

m_c = missed

- Let $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_C)$ = probabilities for $\boldsymbol{n} = (n_1, \ldots, n_C)$ that characterize the distribution of $\boldsymbol{X}$ for the captured persons; can be drawn from the posterior distribution from a multilevel LC model for persons within blocks

$$\boldsymbol{n} \mid n_+, \boldsymbol{\gamma} \sim \text{Mult}(n_+, \boldsymbol{\gamma})$$

- Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_C)$ = capture probabilities drawn from the posterior distribution under P-sample logistic model
- Conditionally given $(\boldsymbol{\gamma}, \boldsymbol{\pi}, N_+)$, the total counts $\boldsymbol{N} = (N_1, \ldots, N_C)$ and the missing counts $\boldsymbol{m} = (m_1, \ldots, m_C)$ are multinomial with cell probabilities proportional to γ_c/π_c and $\gamma_c\ (1 - \pi_c)/\pi_c$, respectively
- Under a diffuse prior for N_+, the posterior distribution for the missing counts given the observed counts is

$$m_+ \mid \boldsymbol{n}, \boldsymbol{\gamma}, \boldsymbol{\pi} \sim \text{NegBin}(\text{size} = n_+, \text{mean} = n_+(1 - \tilde{\pi})/\tilde{\pi})$$

$$\boldsymbol{m} \mid m_+, \boldsymbol{\gamma}, \boldsymbol{\pi} \sim \text{Mult}(m_+, \boldsymbol{\gamma}^*)$$

$$\gamma_c^* \text{ proportional to } \gamma_c\ (1 - \pi_c)/\pi_c$$

United States® Census Bureau

Some Issues Yet to be Addressed

Current round of TUA

- How to incorporate prior information about sex ratios from demographic analysis (In PES, dual-system estimates (DSEs) were adjusted to match sex ratios provided by Census Bureau demographers at the national level, to help correct the DSEs for correlation bias.)
- How to summarize posterior draws. (Likely scenario: Each draw from the posterior distribution of the “true” census will be tabulated and compared to the corresponding tabulations from the CEF; the distribution of these discrepancies over the posterior draws represents the uncertainty.)
- How to release summaries of TUA while controlling disclosure risk

Future rounds of TUA

- Modeling errors in person characteristics using administrative records
- Accounting for additional noise added by Disclosure Avoidance
- Use of administrative records, Demo Frame, and more sophisticated capture-recapture models (e.g., log-linear or latent-class analysis) for better estimates of omissions

United States® Census Bureau

References

Cohen, M.P. (1997) The Bayesian bootstrap and multiple imputation for unequal probability sample designs. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 635-638.

Dong, Q., Elliott, M.R. and Raghunathan, T.E. (2014) A nonparametric method to generate synthetic populations to adjust for complex sample design. *Survey Methodology*, 40:29-46.

Dunson, D.B. and Xing, C. (2008). Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association,* 104:487, 1042-1051.

Gerber, F., Mosinger, K., Furrer, R. (2018). dotCall64: An R package providing an efficient interface to compiled C, C++, and Fortran code supporting long vectors. *SoftwareX*, 7, 217-221.

Hu, J., Reiter, J.P. and Wang, Q. (2018). Dirichlet process mixture models for modeling and generating synthetic versions of nested categorical data. *Bayesian Analysis*, 13:3, 183-200.

Little, R.J.A. and Zheng, H. (2007) The Bayesian approach to the analysis of finite population surveys. *Bayesian Statistics*, 8:283-302.

Manrique-Vallier, D. and Reiter, J.P. (2014a). Bayesian multiple imputation for large-scale categorical data with structural zeros. *Survey Methodology*, 40:1, 125-134.

Manrique-Vallier, D. and Reiter, J.P. (2014b). Bayesian estimation of discrete multivariate latent structure models with structural zeros. *Journal of Computational and Graphical Statistics*, 23:4, 1061-1079.

United States® Census Bureau

Rubin, D.B. and Zaslavsky, A.M. (1989) An overview of representing misenumerations in the Census using multiple imputation. Proceedings of the Bureau of the Census Fifth Annual Research Conference, 109-117.

Vermunt, J.K. (2003). Multilevel latent class models. *Sociological Methodology*, 33:1, 213-239.

Vermunt, J.K. (2008). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research,* 17:1, 33-51.

Vermunt, J.K. and Magidson, J. (2016). *Technical Guide for Latent GOLD 5.1: Basic, Advanced, and Syntax.* Belmont, MA: Statistical Innovations, Inc.

Vermunt, J.K., Van Ginkel, J.R., Van der Ark, L.A. and Sijtsma, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology*, 38:1, 369-397.

Vidotto, D., Kaptein, M.C. and Vermunt, J.K. (2015). Multiple imputation of missing categorical data using latent class models: state of the art. *Psychological Test and Assessment Modeling*, 57:4, 542-576.

Vidotto, D., Vermunt, J.K. and Van Deun, K. (2018a). Bayesian latent-class models for the multiple imputation of categorical data. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 14:2, 56-58.

Vidotto, D., Vermunt, J.K. and Van Deun, K. (2018b). Bayesian multilevel latent-class models for the multiple imputation of nested categorical data. *Journal of Educational and Behavioral Statistics*, 43:5, 511-539.

United States® Census Bureau

Zaslavsky, A.M. (2004) Representing the Census undercount by multiple imputation of households. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* (Eds. A. Gelman and X.L. Meng). West Sussex, England: John Wiley & Sons.

Zanutto, E.L. and Zaslavsky, A.M. (2006). A model for estimating and imputing nonrespondent census households under sampling for nonresponse follow-up. *Survey Methodology*, 32(1), 65-76.

Zhou, H., Elliott, M.R. and Raghunathan, T.E. (2016) Synthetic multiple-imputation procedure for multistage complex samples. *Journal of Official Statistics*, 32(1): 231-256.