Day 2 follow-up

William Sexton

On behalf of and with the support of the 2020 DAS development team
U.S. Census Bureau
CNSTAT
December 12, 2019

Thanks to the 2020 DAS development team and collaborators: Tammy Adams, Robert Ashmead, Craig Corl, Ryan Cummings, Jason Devine, John Fattaleh, Simson Garfinkel, Nathan Goldschlag, Michael Hawes, Michael Hay, Cynthia Hollingsworth, Michael Ikeda, Kyle Irimata, Dan Kifer, Philip Leclerc, Ashwin Machanavajjhala, Christian Martindale, Gerome Miklau, Claudia Molinar, Brett Moran, Ned Porter, Sarah Powazek, Vikram Rao, Chris Rivers, Anne Ross, Ian Schmutte, William Sexton, Rob Sienkiewicz, Tori Velkoff, Lars Vilhuber, Bei Wang, Tommy Wright, Bill Yates, and Pavel Zhurlev.

The views in this presentation are those of the author, and not those of the U.S. Census Bureau.

Shape your future START HERE >

United States® Census 2020

Known Issues

- **There are two sources of error in the TopDown Algorithm (TDA):**
  - Measurement error due to differential privacy noise
  - Post-processing error due to statistical inference creating non-negative integer counts from the noisy measurements
- **Post-processing error tends to be much larger than differential privacy error**
- **Positive bias in small counts/negative bias in large counts is the result of**
  - Invariants
  - Post-processing error specifically introduced by our L2 optimization routine
- **Improving post-processing is not constrained by differential privacy**
- **Techniques to improve post-processing error may be drawn from demography, statistics, computer science, operations research, econometrics, etc. without increasing the privacy-loss budget**

Non-negative Least Squares (NNLS)

- **The post-processing L2 solve (NNLS) finds the best fitting non-negative histograms**
- **The differential privacy measurements constrain this search**
  - Closeness is measured by mean squared error
  - Other constraints include: invariants, structural zeros, hierarchical consistency (tables add up)
  - Measurements include the detailed histogram query and several marginal queries:
    - The detailed histogram permits creating micro-data, a binding requirement inside the 2020 Census production system
    - The marginal queries are the specific table groups in the PL94-171 and DHC specifications; this is how they are made more accurate
- **NNLS is not Ordinary Least Squares (OLS)**
  - If it were, then the solution would be provably minimum variance unbiased (in the class of linear estimators)
  - We are working on a hybrid solution that uses OLS when it can (hence, minimum variance unbiased) and NNLS otherwise
  - This is not a panacea, but will result in accuracy improvements without additional privacy-loss budget
- **Reducing the post-processing error is not a privacy research problem**
- **It is a statistical research problem**
- **It is also the primary research focus of the DAS scientific team**
- **Collaboration is welcome!**

Design of the TDA measurements

- **TDA optimizes for counts not ratios or other non-linear functions**
  - Alternative methods may be required to address use-cases involving ratios or other non-linear functions
  - Those methods will probably work better if they start from the original differentially private measurements
  - Examples include demographic forecasting and spatial segregation models: the plug-in estimator using the official tables is not the optimal statistical estimator
  - Providing direct access to the differentially private measurements does not require the use of the FSRDCs
  - It does require supporting alternative releases (in addition to the official release) of the 2020 Census data
  - Given resource constraints and policy implications of releasing alternative products, we would like to hear from the user community before committing to producing an alternative set of data products
    - Measures of uncertainty are straightforward with the DP measurements used by TDA
    - The measurements exhibit inconsistency, which was the driving force behind the micro-data output requirement
  - And the Census Bureau must have the resources to support them (policy decision)

Geographic allocation of the PLB

- **TDA expends privacy-loss budget on the central hierarchy (a.k.a. spine)**
  - The current TDA has an extra layer (tract groups). The suggestion to use these programmatically is a good one, which we will investigate
  - This design directly supports the redistricting application:
    - Virtually all legislative bodies are within political geographies, which are predominantly county- or state-based
    - We cannot put future districts onto the spine (they are unknown when PL94-171 is produced)
    - The major legislative bodies are on the spine
    - The design ensures that legislative districts will have the most accurate boundaries and VRA determinations
- **TDA does not directly allocate PLB off-spine**
  - Creating separate geographic spines would be a major redesign of TDA (policy decision, not an engineering consideration)
  - School districts, AIAN tribal areas, etc. do not receive a direct share of the privacy-loss budget
  - Research suggests that this design feature may have created unintended consequences including inequities
  - These are being documented and addressed, including tribal consultations to address the AIANNH concerns
  - Adding custom queries that embody important information about certain off-spine geographies is feasible within the current design
    - Introduce special queries that aggregate over combinations of cells with small expected sums
    - Cells selection procedure cannot violate differential privacy
    - Choice can be informed by general knowledge and public information such as past Censuses or the American Community Survey
    - The potential gain in accuracy from choosing well far outweighs the potential loss of choosing poorly

Vacancy Rates

- Vacancy rates in the 2010 Demonstration Data Products often dropped significantly as compared to the original SF-1 (where they were invariant)
- This is a direct, but unintended, consequence of the 2010 Demonstration Data Products design (subset of the full DHC specifications)
- The full DHC includes the additional tabulations and queries required to fix this issue

Allocation of the PLB across tables

- **The full PL94-171 and DHC specifications involve an enormous number of statistics**
  - Approximately 2.5M at each level of the central geographic hierarchy (including tract groups)
  - The current allocations represent best efforts to tune the allocation among these queries (algorithmic and by-hand)
  - Based on the instruction to insure that the redistricting application remains fit-for-use, allocate the balance to other queries
  - Continuing research and collaboration is welcome here, too
- **Defer to the closing discussion policy-based decisions to re-arrange the PLB**

Current Status and Path Forward

- **Re-allocation and re-design are outside the scope of this presentation**
- **Raise those questions during the closing discussion**
- **Feedback is welcome at any time although the sooner the better**
- **The most helpful actionable feedback**
  - Identification of impossible or improbable outcomes in the 2010 Demonstration Data Products
  - Suggestions that could be used to improve the design and optimization of the DAS to produce data products with the highest fitness-for-use
  - Acceptable tradeoffs with results-oriented objectives along the lines of (e.g., “willingness to sacrifice some existing accuracy at the block level to improve tract-level data”) or standards-based thresholds (e.g., “county/tract/block-level data needs to be at least X/Y/Z% accurate to be acceptable”)
  - We want your code, and we will work with you to implement some of these analyses internally