

# Updates and DAS Development Schedule

## **Michael Hawes**

Senior Advisor for Data Access and Privacy  
Research and Methodology Directorate  
U.S. Census Bureau

CNSTAT Expert Meeting on Disclosure Avoidance  
March 18, 2020

Any opinions and viewpoints expressed in this presentation are the author's own, and do not represent the opinions or viewpoints of the U.S. Census Bureau.

Shape  
your future  
START HERE >

United States<sup>®</sup>  
Census  
2020

# Demonstrating Privacy, Assessing and Improving Accuracy

The DAS Team's priorities over Fall 2019 were:

- To scale up the DAS to run on a (nearly) fully-specified national histogram
- To demonstrate that the DAS can effectively protect privacy at scale
- To permit the evaluation and optimization of the DAS for accuracy and “fitness for use”

These initiatives were largely successful, but much more work needs to be done over the remainder of this year.

The engagement and efforts of our data users have been enormously helpful in helping to identify and prioritize this remaining work.

# Committee on National Statistics Workshop

**December 11-12, 2019**

**Evaluation of the Demonstration Data Products (DDP): 2010 Census data run through a preliminary version of the 2020 DAS**

**Data user assessments and findings on DAS implications for:**

- **Redistricting and related legal use cases**
- **Identification of rural and special populations**
- **Geospatial analysis of social/demographic conditions**
- **Delivery of government services**
- **Business and private sector applications**
- **Denominators for rates and baselines for assessments**

# What We've Learned

**The October vintage of the DAS falls short on ensuring “fitness for use” for several priority use cases.**

## **Particular areas of concern:**

- Population counts for political geographies
- Population counts for American Indian and Alaska Native Tribes and Tribal Areas
- Systemic biases (e.g., urban vs. rural)
- Housing statistics and vacancy rates

**These issues are substantially driven by post-processing of the noisy statistics within the DAS.**

# What We've Learned

- **There are two sources of error in the TopDown Algorithm (TDA):**
  - Measurement error due to differential privacy noise (tunable through selection of  $\epsilon$ )
  - Post-processing error due to process of creating internally consistent, non-negative integer counts from the noisy measurements
- **Post-processing error tends to be much larger than DP error**
- **Improving post-processing is not constrained by DP**

# Causes of Post-Processing Error

## Sparsity!

**Earlier runs of the DAS (e.g., 2018 E2E Test) processed a smaller histogram, where most cells were populated.** (2,012 statistics = ~22 Billion cells at the block level)

**The DDP included a much larger histogram.** (400,000 statistics = ~4.4 Trillion cells at the block level)

**The more statistics you calculate, the greater the likelihood of a pull from the tail of the noise distribution.**

**Within the constrained population totals of higher geographic levels of TDA, the algorithm had difficulty prioritizing legitimate positive values against all the “noisy” zeros.**

# Current Initiatives

## **Incorporating legal and political geographies into the geographic hierarchy or “spine”**

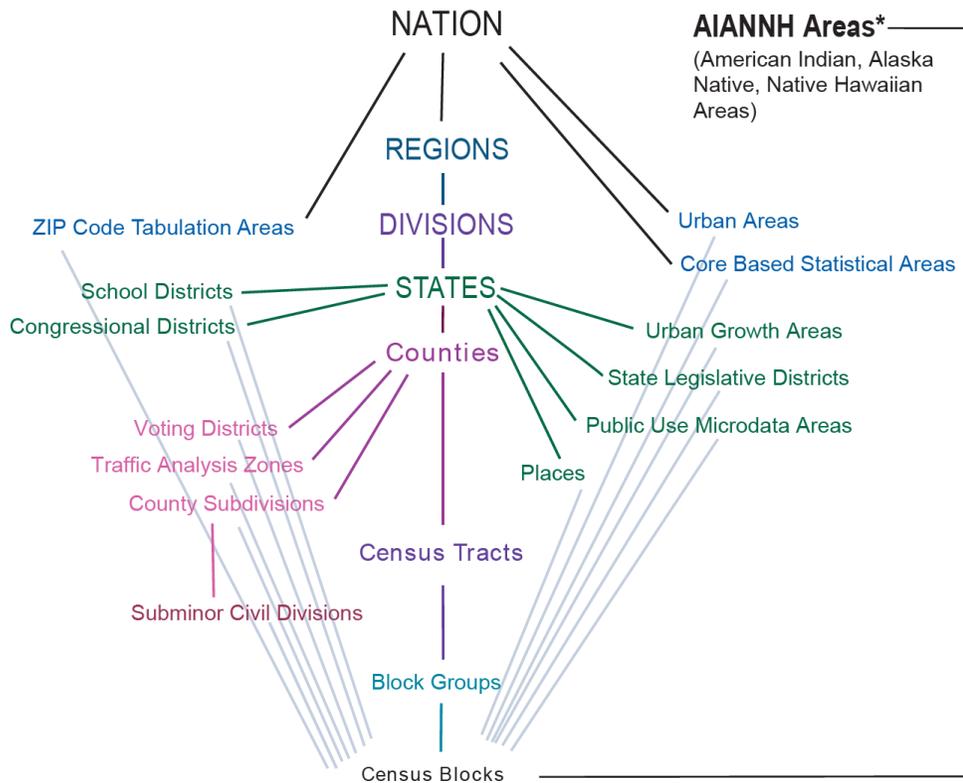
- Allows for direct measurement of statistics of interest (vs. aggregation from block-level data)
- Provides legal and political entities with a dedicated share of the privacy-loss budget.

## **Adopting a multi-phase approach to post-processing**

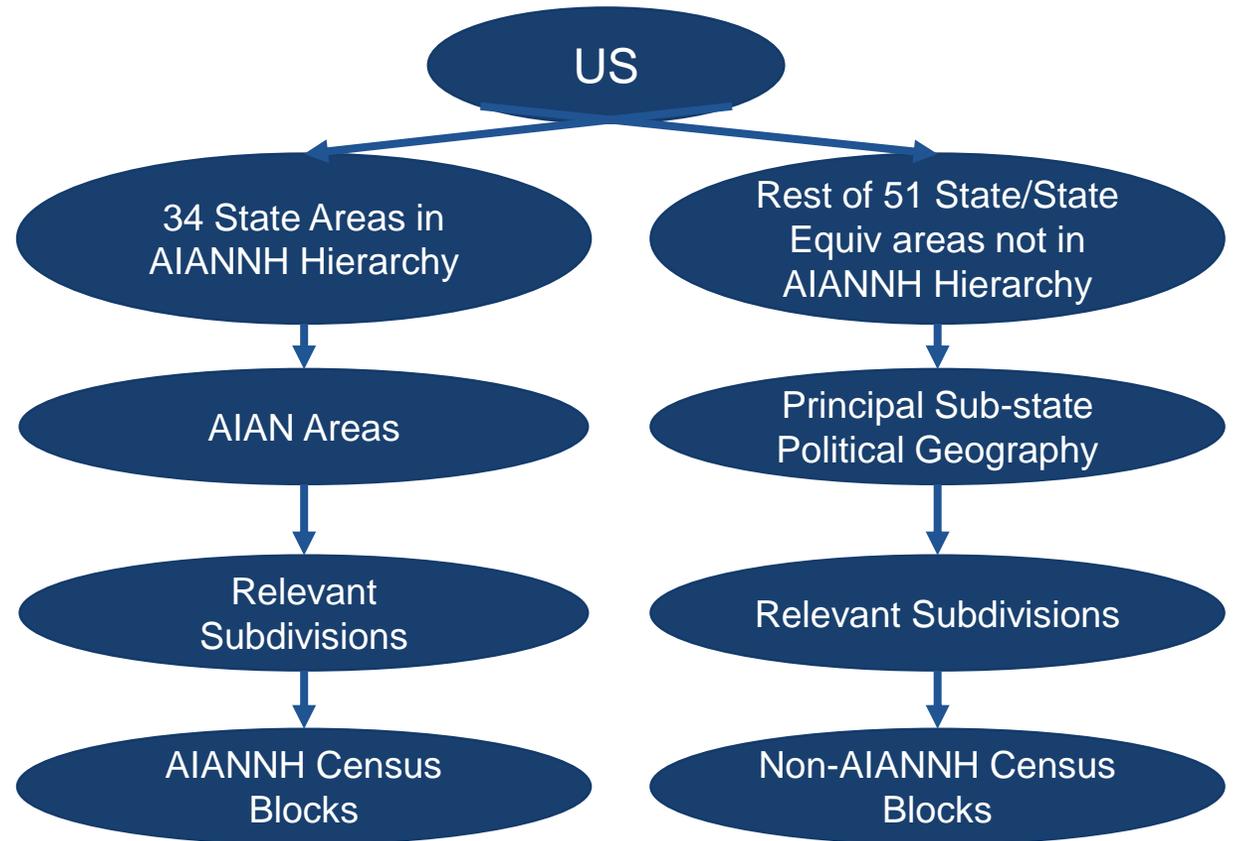
- Addresses the sparsity issue
- Allows for better prioritization of use cases

# Geographic Hierarchy

## Old Hierarchy:



## New Approach (work in progress):



# Multi-Pass Post-Processing

## Old Approach

### Single-pass post-processing:

- **Optimize accuracy for ~1.2M histogram cells** (DDP used only ~400,000 cells)
- **All cells must be integers**
- **All cells must be  $\geq 0$**
- **All margins must satisfy adding up constraints within and between levels of the geographic spine**
- **All invariants and structural zeros must hold exactly**

## New Approach (work in progress)

### Multi-pass post-processing:

- **First pass: compute total population and GQ populations**
- **Second pass for redistricting file** (total pops constrained to first pass values)
- **Third pass for population-estimates program. 3M tabs.** (counts constrained to second pass values)
- **Fourth pass: rest of DHC-H and DHC-P** (counts constrained to values from passes above)

# DAS Development Cycle

Improvements to the DAS are being done using an *Agile* approach.

Series of 4 week development sprints followed by a 2 week evaluation and planning window.

**Current sprint: February 27-March 26**

- Implementing multi-phase processing
- Finalizing details of new geographic hierarchy

**Next sprint begins: April 9**

- Implementing new geographic hierarchy

# Key Dates

**September 2020**

**DSEP finalizes policy decisions on DAS design**  
(e.g., geography, processing, invariants)

**December 2020**

**DSEP sets privacy-loss budget for Group I products**

**January 2021**

**Production run of DAS for Group I products**

**February-March 2021**

**Release of PL94-171 (redistricting files)**

**Summer 2021**

**Release of Demographic Profiles and Demographic and Housing Characteristics files**

# Questions?

## **Michael B. Hawes**

Senior Advisor for Data Access and Privacy  
Research and Methodology Directorate  
U.S. Census Bureau

301-763-1960 (Office)

[michael.b.hawes@census.gov](mailto:michael.b.hawes@census.gov)