# Census Differential Privacy and Private Sector Data Products

Ken Hodges

Sarah Burgoyne

claritas

# The Claritas "Use Case"

- Not a specific question or decision
- Building information products
  - For many businesses
  - Many use cases
- Demographic estimates
  - Build from census data
  - For small areas nationwide
- Concerned with <u>overall</u> impact of Differential Privacy (**DP**)

claritas

# The Claritas Analysis

- Comparing Basic Totals
  - Population, Households, Housing Units, Group Quarters, Family HHs
- Selected characteristics
- Geographic levels
  - Block group
  - Tract
  - County
  - State
- Work still in progress
- Preparing a paper

# The Claritas Analysis

- Need broad measures of difference
  - DP vs. Published 2010
  - How do DP data "behave?"
    - Do small area data sum to large area data?
    - Do differences diminish for larger areas?
    - Do DP data pass consistency checks?
- Census Bureau reminds us
  - DP differences are not necessarily <u>errors</u>
  - Published 2010 also had error
    - Introduced by swapping

claritas

# FINDINGS

# Do Small Area DP Data Sum to Large Area DP Data?

- The quick answer is **"YES"** (as Census Bureau assures)
  - We checked for several tables
    - Block groups summed exactly to Tracts
    - Tracts summed exactly to Counties
    - Counties summed exactly to States

- Important for business applications
  - Improved accuracy with aggregation
  - Example:  Block groups
    - Aggregated to 20 min. drive time around a store

claritas

# BASIC TOTALS

# Basic Totals

- Housing Units:  DP same as published
- All others differ

**Mean** Absolute Percent Difference:  DP vs. Published 2010

| Geog Level | N | Housing Units | Households | Population |
|---|---|---|---|---|
| Block Group | 217,182 | 0.0 | 11.1 | 3.0 |
| Tract | 72,739 | 0.0 | 8.8 | 3.7 |
| County | 3,143 | 0.0 | 9.6 | 0.8 |
| State | 51 | 0.0 | 0.2 | 0.0 |

- Note:  Aggregation does not always reduce mean difference
- Medians smaller (as expected).  Always improve with aggregation
- Interesting outliers

# Basic Totals

- Outliers:  Not just change among very small numbers

| Geog Level | Pop Pub | Pop DP | Diff | Pct Diff |
|---|---|---|---|---|
| 15 003 9808.00 1 | 1 | 69 | 68 | 6,800.0 |
| 06 037 5409.02 4 | 1 | 66 | 65 | 6,500.0 |
| 22 115 9507.01 1 | 4 | 128 | 124 | 3,100.0 |

A closer look at 22 115 9507.01 1

|  | Pop | GQ | HU | HH | PPH |
|---|---|---|---|---|---|
| Pub 2010 | 4 | 0 | 3 | 2 | 2.00 |
| DP 2010 | 128 | 0 | 3 | 3 | 42.67 |

# Basic Totals

- Outliers: Not just change among very small numbers

| Geog Level | Pop Pub | Pop DP | Diff | Pct Diff |
|---|---|---|---|---|
| 31 109 0035.00 1 | 212 | 1 | -211 | -99.5 |
| 49 049 0027.01 4 | 328 | 5 | -323 | -98.5 |
| 06 035 0404.00 1 | 1,296 | 764 | -532 | -41.0 |

A closer look at 06 035 0404.00 1

| | Pop | GQ | HU | HH | PPH |
|---|---|---|---|---|---|
| Pub 2010 | 1,296 | 47 | 579 | 478 | 2.61 |
| DP 2010 | 764 | 2 | 579 | 579 | 1.32 |

# Basic Totals

- Initially more concerned with characteristics
  - Surprised by differences in totals
- IF swapping did not change totals, DP differences are <u>errors</u>
  - Errors built into private sector estimates

<u>Important Because</u>

- Census totals have been standard for accuracy
  - The way we evaluate accuracy of our estimates
  - The way to judge accuracy of private databases
- Will Census totals still be the standard?
- If not . . .
  - How will we evaluate our 2020 estimates?
  - How can we check claims of commercial database providers?

claritas

# CONSISTENCY CHECKS

# Consistency Checks

- Claritas estimates of basic totals
  - Pop, HU, HH, GQ, Fam HHs
- Required to pass consistency checks
  - Check 1:  Households must be less than or equal to Housing Units
  - Check 2:  Family Households must be less than or equal to Households
  - Check 3:  GQ population must be less than or equal to Total Population
  - Check 4:  HH population must be greater than or equal to Family HHs * 2
  - Check 5:  Persons Per Household must be greater than or equal to 1.00
- Published 2010 pass all checks at all levels
- What about DP 2010 data?

claritas

# Consistency Checks

DP Census Data Failing Consistency Checks

| Level | N | Check 1 | Check 2 | Check 3 | Check 4 | Check 5 |
|---|---|---|---|---|---|---|
| Block Group | 217,182 | 0 | 0 | 0 | **1,138** | **313** |
| Tract | 72,739 | 0 | 0 | 0 | **250** | **68** |
| County | 3,143 | 0 | 0 | 0 | **38** | **5** |
| State | 51 | 0 | 0 | 0 | 0 | 0 |

- Not many areas, but logical impossibilities
- We reject and correct Claritas estimates with such inconsistencies
- Interesting outliers

# Consistency Checks:  Outliers on PPH

BG 23 005 0170.02 3 (Cumberland County, ME)

|  | Pop | GQ | HU | HH | HHpop | PPH |
|---|---|---|---|---|---|---|
| Published 2010 | 5 | 0 | 481 | 2 | 5 | 2.50 |
| DP 2010 | 7 | 0 | 481 | 150 | 7 | 0.05 |

BG 06 035 0404.00 2  (Lassen County, CA)

|  | Pop | GQ | HU | HH | HHpop | PPH |
|---|---|---|---|---|---|---|
| Published 2010 | 8,126 | 8,110 | 7 | 7 | 16 | 2.29 |
| DP 2010 | 8,533 | 7,840 | 7 | 7 | 693 | 99.00 |

- – 99.000 is greater than 1.00 – but unrealistic
  - – Many that pass are unrealistic
- DP applied separately to population and households

# Consistency Checks:  PPH Outlier Summed to Tract

BGs in Tract 23 005 0170.02  (Cumberland County, ME)

|  |  | Pop | GQ | HU | HH | HHpop | PPH |
|---|---|---|---|---|---|---|---|
| **BG 1** | Pub 2010 | 2,372 | 709 | 808 | 641 | 1,663 | 2.59 |
|  | DP 2010 | 2,405 | 741 | 808 | 598 | 1,664 | 2.78 |
| **BG2** | Pub 2010 | 1,234 | 0 | 931 | 482 | 1,234 | 2.56 |
|  | DP 2010 | 1,214 | 0 | 931 | 402 | 1,214 | 3.02 |
| **BG3** | Pub 2010 | 5 | 0 | 481 | 2 | 5 | 2.50 |
|  | DP 2010 | 7 | 0 | 481 | 150 | 7 | 0.05 |
| **Sum** | **Pub 2010** | **3,611** | **709** | **2,220** | **1,125** | **2,902** | **2.58** |
|  | **DP 2010** | **3,626** | **741** | **2,220** | **1,150** | **2,885** | **2.51** |

- Aggregation can help
- BG 1 and BG 2 pass the PPH check
- The **Sum** to tract passes, and is close to published

# CHARACTERISTICS

# Characteristics

SF1 Table P5:  Population by Race and Ethnicity

- Not Hispanic White
- Not Hispanic Black or African American
- Not Hispanic American Indian or Alaska Native
- Not Hispanic Asian
- Not Hispanic Native Hawaiian or Other Pacific Islander
- Not Hispanic Other
- Not Hispanic  Two or More Races
- Hispanic White
- Hispanic Black or African American
- Hispanic American Indian or Alaska Native
- Hispanic Asian
- Hispanic Native Hawaiian or Other Pacific Islander
- Hispanic Other
- Hispanic  Two or More Races

claritas

# Characteristics

<u>SF1 Table P12</u>:  Population by Age by Sex

The following Age Categories by Male and Female

| 0-4 | 30-34 | 67-69 |
|---|---|---|
| 5-9 | 35-39 | 70-74 |
| 10-14 | 40-44 | 75-79 |
| 15-17 | 45-49 | 80-84 |
| 18-19 | 50-54 | 85 + |
| 20 | 55-59 | |
| 21 | 60-61 | |
| 22-24 | 62-64 | |
| 25-29 | 65-66 | |

# Characteristics

- <u>SF1 Table P28</u>:  Households by Type by Size
    - Family 2 persons
    - Family 3 persons
    - Family 4 persons
    - Family 5 persons
    - Family 6 persons
    - Family 7 or more persons
    - Nonfamily 1 person
    - Nonfamily 2 persons
    - Nonfamily 3 persons
    - Nonfamily 4 persons
    - Nonfamily 5 persons
    - Nonfamily 6 persons
    - Nonfamily 7 or more persons

claritas

# Characteristics

SF1 Table P25:  Households by Presence of Persons Age 65+

- Collapsed to two categories
  - With a Person age 65+
  - Without a Person Age 65+

claritas

# Characteristics

- How different are DP and Published <u>percent distributions</u>?
- Index of dissimilarity (**IOD**)
- IOD ranges from:
  - 0.0 if identical
  - 100.0 if no similarity
- Interpretation
  - Percent of Persons or Households in DP distribution to shift to another category to make it equal the Published distribution

claritas

# Characteristics

- Mean Index of Dissimilarity by Characteristic and Geographic Level

| Table | Block Group | Tract | County | State |
|---|---|---|---|---|
| P5:  Pop by Race/Hispanic | 3.8 | 2.2 | 1.0 | 0.1 |
| P12:  Pop by Age/Sex | 35.4 | 33.4 | 8.8 | 0.1 |
| P25:  HHs by Person Age 65+ | 8.1 | 5.1 | 2.0 | 0.1 |
| P28:  HHs by Type and Size | 18.0 | 11.5 | 6.7 | 0.3 |

- IODs vary widely by characteristic
- Medians only modestly lower.  Similar pattern
- Distribution of Privacy-loss budget?

claritas

# Characteristics

- For perspective:  How much did ACS differ from Published 2010?
    - HHs by Type and Size:  ACS vs. Published
        – ACS sample data
        – 5-Year Period Estimates 2008-2012
        – Centered on 2010

    <u>Mean IOD</u>:  DP and ACS vs. Published 2010 HHs by Type & Size

| Table | Block Group | Tract | County | State |
|---|---|---|---|---|
| DP:  2010 | 18.0 | 11.5 | 5.7 | 0.3 |
| ACS:  2008-2012 | 18.9 | 11.1 | 4.6 | 2.0 |

- Is it OK that DP differs from census as much as ACS differs from census?

# Characteristics:  Race/Hispanic Outliers

| BG 15 003 0110.00 3   IOD = 96.2 | Pub | DP |
|---|---|---|
| Population | 395 | 276 |
| Pct Not Hispanic White | **28.6** | 0.0 |
| Pct Not Hispanic Black | 4.1 | 0.0 |
| Pct Not Hispanic Am Indian | 0.5 | 0.0 |
| Pct Not Hispanic Asian | **17.7** | 0.0 |
| Pct Not Hispanic NHOPI | **39.0** | 0.0 |
| Pct Not Hispanic Other | 0.3 | 0.0 |
| Pct Not Hispanic 2+ Races | 3.8 | **47.5** |
| Pct Hispanic White | 2.5 | 0.0 |
| Pct Hispanic Black | 0.0 | 0.0 |
| Pct Hispanic Am Indian | 0.0 | 0.0 |
| Pct Hispanic Asian | 1.0 | 0.0 |
| Pct Hispanic NHOPI | 1.3 | 0.0 |
| Pct Hispanic Other | 1.3 | 0.0 |
| Pct Hispanic 2+ Races | 0.0 | **52.5** |

claritas

# Characteristics:  Race/Hispanic Outliers

| BG 04 021 0020.02 1  IOD = 87.1 | Pub | DP |
|---|---|---|
| Population | 651 | 1,162 |
| Pct Not Hispanic White | **38.6** | 0.0 |
| Pct Not Hispanic Black | 5.8 | 0.0 |
| Pct Not Hispanic Am Indian | 1.4 | 0.0 |
| Pct Not Hispanic Asian | 0.3 | **31.6** |
| Pct Not Hispanic NHOPI | 0.2 | **55.9** |
| Pct Not Hispanic Other | 0.2 | 0.0 |
| Pct Not Hispanic 2+ Races | 3.5 | 0.0 |
| Pct Hispanic White | **25.8** | 0.0 |
| Pct Hispanic Black | 0.6 | 0.0 |
| Pct Hispanic Am Indian | 2.8 | 0.0 |
| Pct Hispanic Asian | 0.0 | 0.0 |
| Pct Hispanic NHOPI | 0.0 | 0.0 |
| Pct Hispanic Other | **18.7** | **12.5** |
| Pct Hispanic 2+ Races | 2.2 | 0.0 |

# Race/Hispanic Outliers (BG 15 003 0110.00 3) Summed to Tract

| BG    15 003 0110.00 3 | BG Pub | BG DP | Tr Pub | Tr DP |
|---|---|---|---|---|
| **Population** | **395** | **276** | **4151** | **4116** |
| Pct Not Hispanic White | **28.6** | 0.0 | **35.3** | **34.3** |
| Pct Not Hispanic Black | 4.1 | 0.0 | 0.7 | 0.3 |
| Pct Not Hispanic Am Indian | 0.5 | 0.0 | 0.2 | 0.0 |
| Pct Not Hispanic Asian | 17.7 | 0.0 | **26.3** | **25.0** |
| Pct Not Hispanic NHOPI | **39.0** | 0.0 | 9.8 | 9.8 |
| Pct Not Hispanic Other | 0.3 | 0.0 | 0.2 | 0.0 |
| Pct Not Hispanic 2+ Races | 3.8 | **47.5** | **21.7** | **22.1** |
| Pct Hispanic White | 2.5 | 0.0 | 2.3 | 1.5 |
| Pct Hispanic Black | 0.0 | 0.0 | 0.0 | 0.0 |
| Pct Hispanic Am Indian | 0.0 | 0.0 | 0.1 | 0.1 |
| Pct Hispanic Asian | 1.0 | 0.0 | 0.6 | 1.1 |
| Pct Hispanic NHOPI | 1.3 | 0.0 | 0.2 | 0.2 |
| Pct Hispanic Other | 1.3 | 0.0 | 0.5 | 0.5 |
| Pct Hispanic 2+ Races | 0.0 | **52.5** | 2.1 | 5.3 |
| **Index of Dissimilarity** | | 96.20 | | 4.05 |

# Concluding Remarks

- Demonstration data show impact of DP
  - Some findings are unsettling

Differences in basic totals
  - Sometimes large (and suspect)
  - Regarded as errors
- Differences not consistent across counts
  - Unrealistic, sometimes impossible, values of PPH

claritas

# Concluding Remarks

Differences in characteristics

- Vary widely by characteristic
- Aggregation helps, but not always

- Differences not necessarily errors
  - Swapping also infuses noise
  - But published 2010 (with swapping)
    - The best standard **WE** have
  - Published 2010 a reasonable standard
    - Seen as providing insufficient protection   (not enough noise)
    - Likely more accurate than DP
    - ALSO:  Some DP data strain credibility

claritas

# Concluding Remarks

Private Sector Priorities

- Biggest concern is with basic totals
  - Do they have to be that different?
  - Can we make them pass consistency checks?

- For characteristics – focus on the basics
  - Age/sex (5 year age breaks)
  - Basic race/Hispanic categories (don't every combination)

- We understand the challenges Census Bureau faces
  - Want to remain strong advocates of the census
  - Look forward to staying engaged as 2020 products are developed

claritas

# Thank You