Feedback on 2020 Census products proposals and 2010 Demonstration products

Cornell University

Program on Applied Demographics

February 2020

Cornell Program on Applied Demographics

Web: https://pad.human.cornell.edu Email: PADinfo@cornell.edu

Introduction

Every decade the suite of products released after the Census changes. Tables are added, deleted, adjusted or the lowest level of geography is changing. Another ongoing change is in techniques applied to avoid disclosure of private information. These changes are driven by changing data needs, changing questionnaires and also advances in technology.

At several moments the Census Bureau asked for feedback on the proposed changes and this document is a collection of the feedback we, the Cornell Program on Applied Demographics, provided in answer to these asks.

Data products

An overview of 2010 data products can be found at:

https://www.census.gov/population/www/cen2010/glance/

The Summary File 1 (SF1) is the most used of these products throughout the decade for uses in the general population.

In 2020, the SF1 will be replaced by the Demographic and Housing Characteristics File (DHC) as the product which will contain most of the data for every day users (Devine, 2019).

Disclosure Avoidance Systems

The Census Bureau applies disclosure avoidance techniques to its publicly released statistical products in order to protect the confidentiality of its respondents and their data. The Census Bureau is bound by title 13 of the US codes, which states that it is prohibited to “make any publication whereby the data furnished by any particular establishment or individual under this title can be identified”.

Different techniques disclosure avoidance were used in the past. (McKenna, 2018)

The 2010 data products encompass more than a billion data cells and the Census Bureau reports that internal research shows that it was able to reidentify 45% of the population using external available data and published Census data.

Differential privacy

Going into 2020 the Census Bureau reconsidered their approach to Disclosure Avoidance and decided it wanted to reduce the number of data cells and apply a mathematical approach to privacy protection. The chosen mathematical approach is called Differential Privacy. (Hawes, 2019)

Much more information on Census 2020 Data products, Disclosure Avoidance and Differential Privacy can be found at:

https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products.html

and

https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html

Opportunities for Feedback

Both a reduction of data cells and a more stringent Disclosure Avoidance System can have an impact on accuracy and usability of the Census data. During several stages of the decision process surrounding the 2020 data products the Census Bureau solicited feedback from data users and other stakeholders.

This report is a collection of the feedback we, here at PAD provided to the Census Bureau to three of those stages:

- Federal Register Notice, asking for feedback on data use and geographic detail
- Crosswalk with proposed data products and shells, asking for feedback on an overview of proposed table changes from the 2010 data products
- 2010 Demonstration products. Asking for feedback on a set of demonstration data that let data users analyze 2010 data with a Differential Privacy based Disclosure Avoidance System applied.

The Census Bureau opened a special email address for feedback: dcmd.2010.demonstration.data.products@census.gov. During the summer/fall important decisions will be made with regards to products and algorithms, but until than the Census Bureau welcomes all feedback.

Federal Register Notice

On July 19, 2018 a Federal Register Notice was published in which the Census Bureau asked about how Census data is being used. (Bureau of the Census, Department of Commerce, 2018)

On October 9, 2018 the period for feedback was extended as announced in a follow-up Federal Register Notice (Bureau of the Census, Department of Commerce, 2018)

From the first FRN:

The Census Bureau is especially interested in receiving responses to the following questions:

1. How are the data from each individual table and data product used? Include any specific legal, statutory, or programmatic uses. Please cite any supporting federal laws or regulations.
2. Why are decennial census statistics used for this purpose? Please provide a clear justification.
3. Without decennial census data, how would this activity be accomplished (e.g., other data sources)?
4. Who are the users of the specific table or data product?
5. Who is affected by the use of the data in this specific table or data product?
6. How much funding is distributed based on these data?
7. What is the lowest level of geography (e.g., county, census block, etc.) at which data need to be published for each specific table?
8. In what additional levels of geography (e.g., county subdivision, school district, etc.) or geographic components (e.g., urban, rural, etc.) do data need to be published for each specific table? If the level of geography specified in the response to item seven relates to the use planned for the levels of geography requested in this response, please explain how they are related.
9. What programmatic, statutory, or legal uses are there for decennial census data that are not being met by the current suite of decennial census products?

A downloadable spreadsheet contains a listing of the data products and specific tables as well as space for feedback: https://www2.census.gov/about/policies/2020-Census-Data-Products-Feedback-Spreadsheet.xlsx. This spreadsheet may be a helpful tool for respondents to provide the requested information, but its use is not required.

Appendix A contains the feedback we sent to the Census Bureau regarding this Federal Register Notice. Feedback on this Federal Register Notice was taken into consideration for the next step: a proposed set of 2020 data products.

Crosswalk with proposed 2020 data products

In September 2019, the Census Bureau released an Excel workbook with proposed data tables for the 2020 data products. For each of the proposed tables was a table number for the Census 2010 data products, making a comparison 'easier'. Also indicated were the lowest level of geography for which each table was published in 2010 and is proposed to be published in 2020.

A slighty updated version of the crosswalk table can be found at:

https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/2020-census-data-products-planning-crosswalk.xlsx

The proposal eliminates many tables and many of the remaining table will not available at the same level of geography. (Vink, Proposed data products, what tables will be available , 2019)

Appendix B contains our feedback on this proposal.

2010 Demonstration data products

In October 2019, the Census Bureau released a set of 2010 Data Products. From their web site (Bureau of the Census, Department of Commerce, 2019):

"To help data users understand how differential privacy may or may not impact data products they are used to receiving, the Census Bureau created demonstration data products for review. This set of data products demonstrate the current computational capabilities of the 2020 Disclosure Avoidance System (DAS). The products include the 2010 Demonstration Public Law 94-171 (P.L. 94-171) Redistricting Data Summary File and the Demonstration Demographic and Housing Characteristics Summary File.

We encourage data users and data scientists to examine the products and provide feedback as we continue to develop and fine-tune disclosure avoidance systems. We are releasing these products to encourage independent analyses from the data user community and a robust, open, data-driven dialogue."

The National Academy of Sciences Committee on National Statistics (CNSTAT) hosted a special workshop about these products on December 11-12, 2019 (CNSTAT, 2019).

Jan Vink (Cornell Program on Applied Demographics) presented here with results from analyses on data on New York School Districts (Vink, Elementary School Enrollment, 2019)

The 2010 Demonstration Data can be downloaded from:

https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/

The Cornell Institute for Social and Economic Research (CISER) reformatted the data as published by the Census Bureau in formats that make it easier for analysts to work with the data (Arguillas, 2019):

https://ciser.cornell.edu/data/data-archive/census-2010-dhc-download-center/

Appendix C contains our feedback on the 2010 Demonstration Products. We analyzed a number of use cases that represent a variety of stakeholders and learned that many stakeholders would have reached different conclusions with the 2010 Demonstration Products than with the published SF1 data. There are also problems with bias and validity.

References

Arguillas, F. (2019). *Census 2010 DHC Download Center*. doi:10.6077/fe6a-f789: https://ciser.cornell.edu/data/data-archive/census-2010-dhc-download-center/

Bureau of the Census, Department of Commerce. (2018, July 19). *Soliciting Feedback From Users on 2020 Census Data Products.* Retrieved from 83 FR 34111: https://www.federalregister.gov/documents/2018/07/19/2018-15458/soliciting-feedback-from-users-on-2020-census-data-products

Bureau of the Census, Department of Commerce. (2018, October 10). *Soliciting Feedback From Users on 2020 Census Data Products; Reopening of Comment Period.* Retrieved from 83 FR 50636: https://www.federalregister.gov/documents/2018/10/09/2018-21837/soliciting-feedback-from-users-on-2020-census-data-products-reopening-of-comment-period

Bureau of the Census, Department of Commerce. (2019). *2010 Demonstration Data Products*. Retrieved from 2020 Decennial Census Program Resources: https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products.html

CNSTAT. (2019, December). *Workshop on 2020 Census Data Products: Data Needs and Privacy Considerations*. Retrieved from Committee on National Statistics, National Academy of Sciences: https://sites.nationalacademies.org/DBASSE/CNSTAT/DBASSE_196518

Devine, J. (2019, November). *Status Update on 2020 Census Data Products Plan.* Retrieved from 2020 Census Data Products: https://www2.census.gov/cac/nac/meetings/2019-11/devine-hollingsworth-status-update-2020-data-products-plan.pdf

Hawes, M. (2019, November). *Title 13, Differential Privacy, and the 2020 Decennial Census.* Retrieved from https://www2.census.gov/about/policies/2019-11-paper-differential-privacy.pdf

McKenna, L. (2018, October). *Disclosure Avoidance Techniques Used for the 1970 through 2010.* Retrieved from Census Bureau Working Papers: https://www.census.gov/content/census/en/library/working-papers/2018/adrm/cdar2018-01.html

Vink, J. (2019, December). *Elementary School Enrollment.* Retrieved from CNSTAT Workshop on 2020 Census Data Products: Data Needs and Privacy Considerations: https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_197511.pdf

Vink, J. (2019, November). *Proposed data products, what tables will be available .* Retrieved from State Data Center Presentations and webinars: https://sdcclearinghouse.files.wordpress.com/2019/12/dp-webinar-vink.pdf

Appendix A: Feedback on Federal Register Notice

Distinction between Group Quarters Population and Household population

Drivers for population change are very different for household populations and for Group Quarters populations. They also require very different levels of support from local communities.

For:

- Local planning
- Population estimates and projections
- Research into changing demographic compositions

It is important to be able to separately understand the composition of the household population and the group quarters population.

Available data in 2010:

	#Sex categories	# Age categories	#Race/Ethnicity categories	#GQ types	Lowest level geography
PCO1 through PCO10 (SF2)	2	18	Very detailed (threshold > 100)	7	County
PCT13, PCT13A-PCT13I (SF1)	2	23	9	Household population only	Tract
PCT20, PCT20A-PCT20I (SF1)	-	-	9	24	Tract
PCT21 (SF1)	2	3	-	24	Tract
PCT39 (SF2)	2	3	Very detailed (threshold > 100)	24	Tract
PCT22,PCT22A-PCT22I (SF1)	2	- (18+)	9	7	Tract
P16, P16A-P16I	-	2	9	Household population only	Block
P42 (SF1)	-	-	-	7	Block
P43 (SF1)	2	3	-	7	Block
QT-P13 (SF1)	2	3	-	2 (by sex/age), 7 (by sex)	Block
GCT-P6 (SF2)	-	-	Very detailed (threshold > 100)	7	MCD, Place, School district, tract

For the kinds of analyses I am referring to, I would like to see:

At the County level: similar to PC01 through PCO10

PCT tables:	a table with full GQ type detail (comparable to the 24 in 2010) 7 main GQ types by major race groups 7 main GQ types by sex and select age groups (0-17, 18-24, 25-34, 35-49, 50-64, 65+) Household Population by age and sex (as in PCT13)
At the block level:	similar to table P43 Household population by sex and select age groups (0-17, 18-24, 25-34, 35-49, 50-64, 65+)

Summary levels for PCT and HCT tables

The 2010 PCT and HCT tables were an important aggregate for many communities as they contained more detail than the block level P tables. It is important to keep those tables available for all tracts AND for areas with a general or special purpose government. Also PCT tables for the areas that represent parts of one governmental area within another (e.g. place within county subdivision – SUMLEV 070) should be preserved.

To understand how the population in an area has changed and for planning purposes tables PCT1, PCT2, PCT11, PCT12, PCT13, PCT14, PCT16, PCT20, PCT21, PCT23, PCT24, HTC1 and HCT4 are especially important to repeat.

Occupancy rate, tenure and household size

By definition the population is equal to number of housing units times occupancy rate time persons per household plus the GQ population.

The ACS is controlled to population estimates and housing units in such a way that this equation holds. A combination of estimates errors and sampling results can cause some very suspect ACS values for occupancies and persons per household and makes the ACS unusable as an alternative.

Trends in occupancy rates (by tenure) and persons per household are very important for planning at very detailed geographic level and the Decennial Census should provide those numbers.

The reason for vacancy is essential to understand the occupancy rates.

Tables involved at block level:

H3,H5,H11,H11A-H11I, H12,H12A-H12I,H13,H14

For planning and research purposes it is also important to look at household size by race of householder and tenure and presence of children. This would require tables at least the tract level.

Tables involved at tract level:

HCT3 (SF2), HCT4(SF2), HCT6(SF2), HCT7 (SF2), HCT12 (SF2)

Headship rates by age by tenure are essential for projecting housing need based on a given set of population projections. This is most important at MCD and tract level.

Tables involved at MCD and tract level:

H17(SF1), HCT8(SF2)

Living situations for elderly

There are many tables that have information on living situations for 65 and older, but do not have any further age breakdown.

With a large part of the baby boom being of the age 65 through 74 it will be harder to draw conclusions about living conditions of the elderly as these change with age and will be biased by the large proportion of the youngest 'old'.

I would like to see information that was presented for 65 and over in 2010 for those that are 65-74 and for 75 and over in 2020.

This information is essential for planning of needs.

Geography should be at least tract level.

Tables involved at tract level:

P34,P34A-P34I, PCT24, PCT28(SF2)

Appendix B: Feedback on Crosswalk with proposed 2020 data products

Geographic summary levels

- PCT, HCT tables
    - In 2010 PCT tables were available for Tracts, Places, MCD, Legislative Districts, ZCTA, School Districts and others.
    - I would like to see a similar list of geographies in the 2020 DHC.
    - If universe thresholds are necessary, I would suggest looking at the tract thresholds and base the population and housing thresholds on that. I suggest a threshold of 1,200 persons for a PCT table and a threshold of 480 housing units for a HCT table.
- PCO tables
    - In 2010 PCO tables were available for Counties and Legislative Districts
    - For 2020 I suggest to produce these tables for all counties and legislative districts (no thresholds), but also for PUMAs (Public Use Microdata Areas), and larger places and larger urban areas. As a threshold for places and urban areas I suggest looking at the ACS 1-year data (65,000) or PUMAs (100,000).
      Motivation: the number of PCO tables is proposed to increase from around 20 in 2010 to around 100 in 2020, making this the geographic summary level with the most tables. If these tables are not available below county level, many of the larger counties are not able to compare areas within the county with each other, e.g. comparing a city with the remainder of the county.

Table recommendations by subject

- Race
    - For ease of use I would like to see the detailed race tallies from PL94 repeated in the DHC so that DHC users don't have to work with two data products
- Age and age by race/ethnicity
    - Single years of age
        - In 2010 we were able to see the population by age and race in PCT tables (PCT12A-PCT12O). The proposal eliminates all cross tabulations between single years of age and race. For population estimates and projections having a good base population is essential and these tables are instrumental.
        - I suggest PCO tables with single years of age for the major race groups. For use as base populations, the H through O tables are more important than A through H as the H through O tables add up to the total population.
    - 23 age groups
        - At some geographic level it is important to have the Non Hispanic population broken out by race and age. This can be accomplished by expanding the proposed P7 subtables from A through I to A through O
    - Age of population in households (and thus in Group Quarters)
        - In 2010 this data was in PCT tables, in 2020 in PCO tables. I think that, because of often significant differences between household and group quarters population, these tables should be available below the county level in large counties. This can be done by publishing PCO tables at more geographic levels (see above) or remain publishing them as PCT tables (or at least the total household population by age)

- Households
    - Population in households and average household size
        - Important indicator that has many applications within estimates and projections. It is important for applying a housing unit method which is often used when information on number of houses is available, and decision makers want to estimate the population in those houses.
        - Having average household size by tenure can easily relay essential information about differences between and within areas. Population estimates based on a housing unit method also benefit from having average household size by tenure.
        - Tables on household size are part of the proposed plans, adding the average household size adds just a little information on the 7+ households.
        - Average household size can be calculated if population in occupied housing units and number of occupied housing units are both available. In the proposal tables on population in occupied housing units by tenure are being eliminated. I think population in occupied housing units by tenure should be available at the block level, as it was in 2010.
    - Household type
        - Household type for the population under 18 years
The proposal eliminates tables P31 [Household type by relationship for the population under 18 years] and P33 [Household type for the population under 18 years in households (excluding householders, spouses, and unmarried partners)]
It is very important to get counts of children living in different types of households. Instead of eliminating I suggest to at least keep table P33 and publish the P31 tables as PCO tables (comparable with the 2010 P34 tables that have such information for the 65plus population). It is not only important to know how many married couple households there are with children, it is also important to know how many children live in these kinds of households. P32 is proposed to remain as PCO table, but does not have the A through I subtables.
        - Household type by race of householder
Was published as P18A through P18I in 2010 and is proposed to be continued as PCO tables. This is an important table to have available below county level for all kinds of decision making at a neighborhood level. I propose to make it PCT tables or at least publish PCO tables for some subcounty geographies.
- Group Quarters
    - In the proposal table PCT22 from SF1 will be published as PST2 in 2020 DHC and the iterations by major race type PCT22A through PCT22I are eliminated in the proposal. The information in PCT22 is already in P15 and there is no reason to bump it up to a PST table.
- Housing Units
    - As argued earlier. Household population by tenure is important for many planning decisions and I would like to see that reinstated (tables H11, H11A through H11I in the SF1). If having these tables at the block level is too problematic, consider publish them as HCT tables.

Appendix C: Feedback on 2010 Demonstration Products

Table of Contents

Introduction

This document analyses data from the Demographic and Housing Characteristics file from the 2010 Demonstration Data Products (DHC).

Much of the document tries to gauge an effect on the quality of the data. The Census Bureau describes as it mission “The Census Bureau’s mission is to serve as the nation’s leading provider of quality data about its people and economy”. One of the pillars of quality data is accuracy and reliability.

Statistics Canada describes accuracy and reliability as[1]:

- **Accuracy** refers to the extent to which the data correctly describes the phenomenon they are supposed to measure.
- **Reliability** is the extent to which the data are accurate consistently over time.
- Accuracy is often decomposed into
  - **precision**, which measures how similar are repeated measurements of the same thing, and
  - **bias**, which measures any systematic departures from reality in the data.
- Other factors contributing to accuracy and reliability are
  - **validity**, the extent to which variables in the dataset have values that correspond to expected outcomes, and
  - **consistency**, the extent to which the data are free of contradiction.

This document will first give some examples of problems with consistency and validity and then define some use cases that look at differences between the published SF1 and the DHC to gauge what difference the Disclosure Avoidance System (DAS) made to the other aspects of accuracy and reliability.

The analyses are performed on New York data retrieved from the Cornell DHC download Center[2]. SAS and Stata code, as well as spreadsheets are available on request.

Conclusions and recommendations

Conclusions

- Treating the person tables and housing tables independently within the DAS leads to inconsistencies within the data and invalid results
- The DAS introduces many more zeroes in the histograms which adds to problems with validity
- There is spatial and temporal autocorrelation in population data:
  - blocks have likely more in common with neighboring blocks than blocks farther away and
  - presence and demographic behavior of most birth cohorts close to each other are correlated.

  These autocorrelations are not taken into account in the DAS which leads to much more randomness in the data, whether we look at age compositions or do spatial analyses

[1] Data quality toolkit: https://www.statcan.gc.ca/eng/data-quality-toolkit (retrieved 1/9/2020)

[2] Census 2010 DHC Download Center https://ciser.cornell.edu/data/data-archive/census-2010-dhc-download-center/

- Many use cases show that decision makers will be faced with data that they know do not describe the situation they need to make decisions on. This knowledge can be based on extreme values, expertise and/or available administrative data
- If the DHC data is all decision makers have, many conclusions and decisions will be different compared to conclusions and decision based on SF1 data
- The DAS tends to bring higher counts down and lower counts up, which introduces biases. For example, sparsely populated areas in New York get more population, more diversity and less aged.

Recommendations

- Enhance the DAS algorithm such that inconsistencies between housing tables and person tables are eliminated
- Increase the privacy budget by a lot: the costs of wrong conclusions and decisions are enormous as there are many big and small entities making decisions which add up quickly
- Adjust the algorithms such that it results in a large decrease in the number of zeroes in the histograms. As I understand it, negative counts go into the post-processing and are in the objective function; this would cause a possible large cost to the objective function if that count goes from zero to one and the optimization model will never adjust that zero
- The last step of Count Review (23-3.5) has as purpose "Prior to releasing census data for external use, conduct a review of final counts after tabulation recodes of the demographic data are applied and the application of disclosure avoidance procedures are performed **to ensure that changes in the counts at multiple levels of geography are reasonable**. This review also ensures that issues identified in the DRF-2, CUF, and CEF reviews have been addressed."[3]
    - All parties at the Census Bureau (and maybe also some experts outside of the Census Bureau) should agree ahead of time what changes in the counts are considered reasonable and not reasonable. Feedback received on the demonstration products might help guide this process.
        - Based on our own analyses and the analyses of others I seriously wonder if this demonstration product would have passed this Count Review step
    - This process only works if there is a mechanism to deal with failed reasonability checks. It is my understanding that rerunning the DAS is very time-consuming, so make sure it can deal with last minute negative findings

[3] Count Review Operation (CRO): Business Process Model: https://www2.census.gov/programs-surveys/decennial/2020/program-management/planning-docs/CRO-bpm.pdf

Consistency

Examples of inconsistencies in the data:

- Inconsistencies between household population (P16) and households by HH type (P18) and between household population and households by HH size (H13)
    - Examples (Universe: blocks in NY):
        - 26,684 blocks with a household population < number of householders
            - 13,573 blocks have zero household population, but do have households
        - 4,077 blocks have zero householders, but have a household population
        - 15,899 blocks with only HH type "non-family households living alone" but with an average HH size > 1
        - 216,852 blocks with no households with 7 or more persons where household population from table P16 does not equal household population calculated from table H13
        - 0 blocks with only family households (per definition 2 or more related people), but with households size 1
- Inconsistencies between Sex (P12) and Household Type (P18)
    - 31,289 blocks with either only men or only women, but with 1 or more married couples

Validity

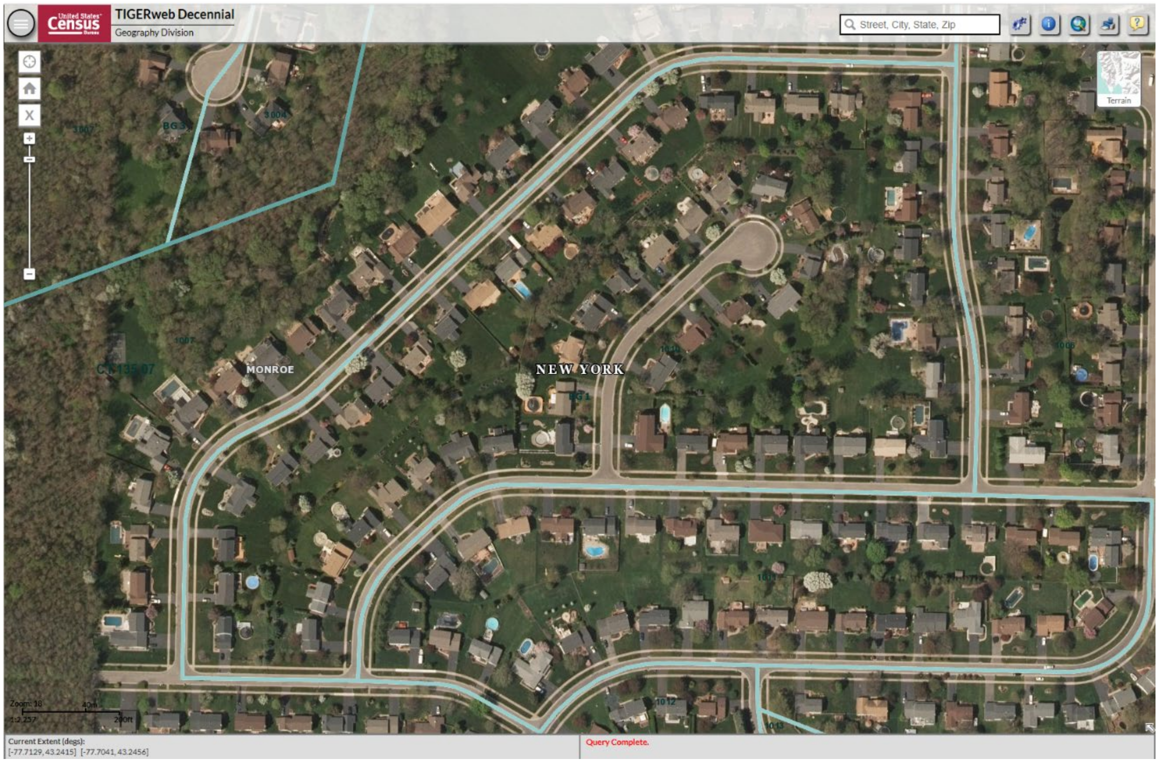
Improbable results (Credibility). Almost all of the measures we looked at in the use cases had a range of outcomes in the Demonstration data that was much larger than the range of outcomes in the SF1. And often the minimum and/or maximum values were outside an expected range of outcomes. This was the case for median age, sex ratio's, women to child ratio's, ratio's with administrative data, etc. Here we will present a few examples of unexpected outcomes (invalid), that will hurt the credibility of the Census results.

- *Blocks with all residents 14 years or younger* (Universe: blocks with only household population and more than 20 residents)
    - No such blocks in SF1, there was 1 block where the oldest age group observed was 18-19 yr old
    - 319 blocks in the demonstration data with all residents 14 yr or younger
        - Example block 360191002004024 had 24 boys age 5-9 and 34 girls age 5-9 and no other age groups present

  From the same analyses:
    - In SF1 94% of blocks in this Universe have at least 1 person 70 year or older
    - In the demonstration data only 54% of the blocks have at least 1 person 70 yr or older

- *Blocks with only males or only females* (blocks with only household population)
    - In SF1 8 blocks with 25 or more males or females and none of the other sex; the largest ones were blocks with people living in special living arrangements (total 778 persons live in a block with no opposite sex and at least 24 other persons of the same sex)
    - In the Demonstration data 9,220 blocks had 25 or more males or females and none of the other sex (a total 356,372 persons live in a block with no opposite sex and at least 24 other persons of the same sex)

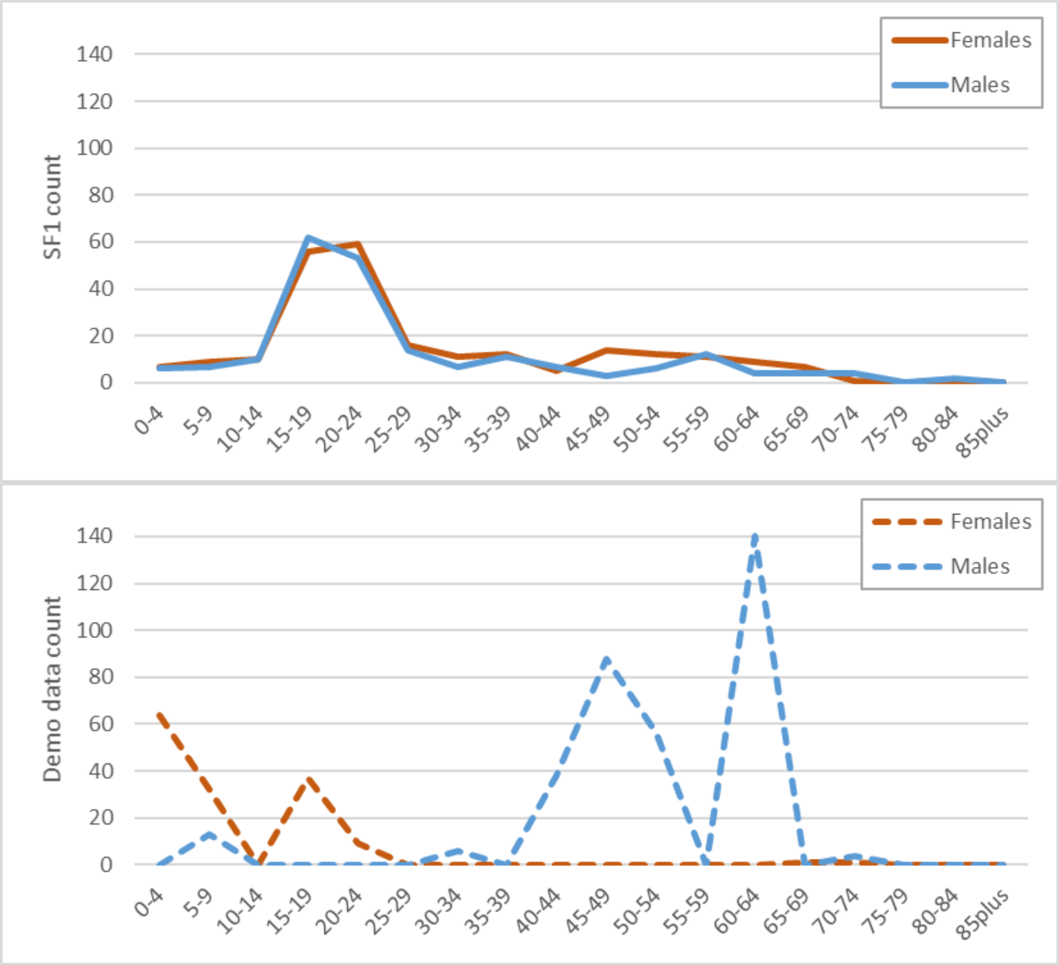
Example: Block 360550135071010 in Rochester city, NY



According to the demonstration data, there were 163 females residing in 59 occupied housing units in this block, no males. In the SF1 data, there were 81 males, 94 females.

- *Age distribution by sex and race*

  There is a strong connection between type of county and age/sex distribution and sometimes a single glance at a population pyramid says something about the type of county (aged community, college county, prison county, etc.). This connection gets lost in the demonstration data. Example:



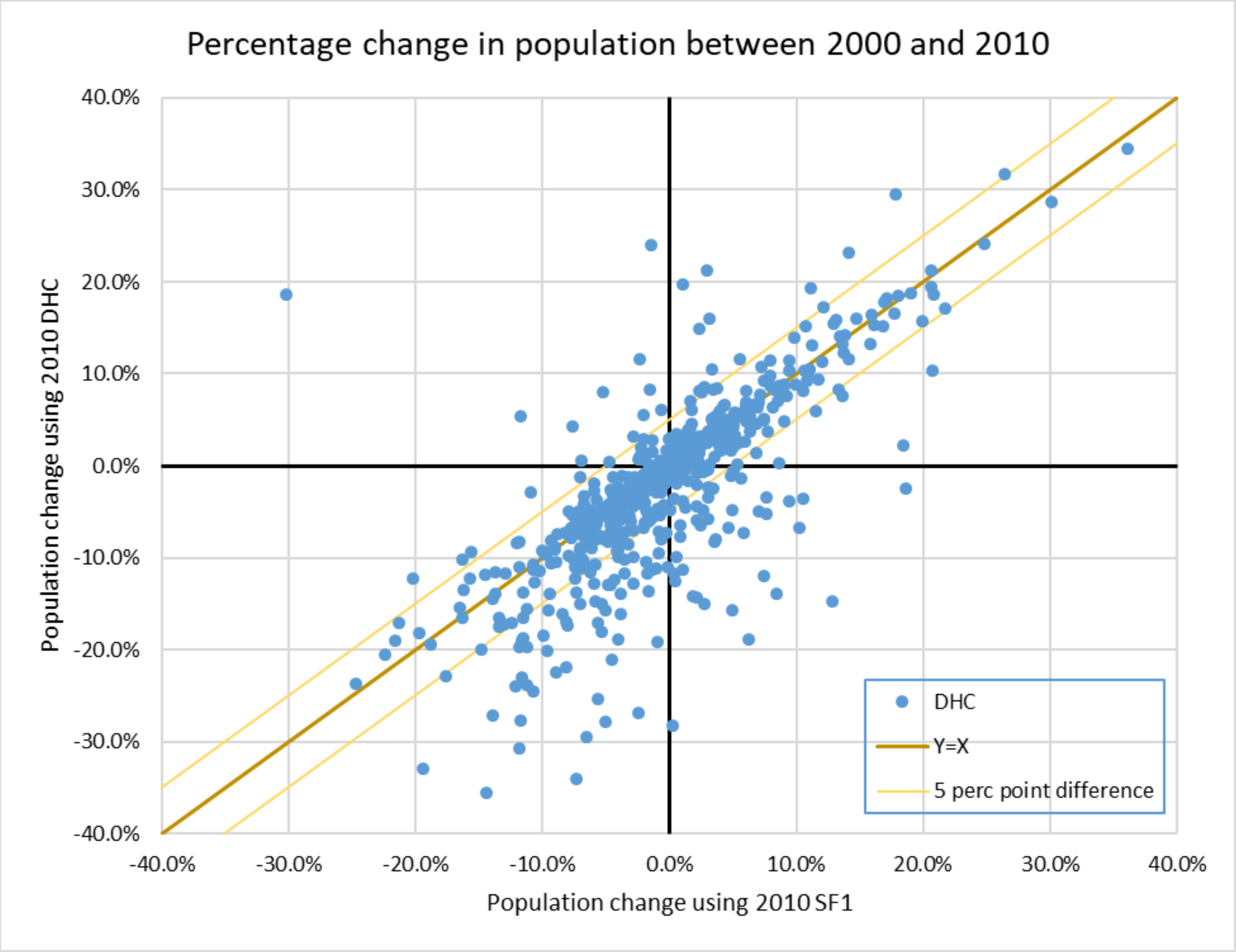
The chart shows the age composition of the Asian population in Allegany county. The SF1 data displays the connection between the presence of a college there and the age distribution. There is no longer that connection in the Demonstration data and furthermore it shows only female 0-4 yr olds and only male 25-69. No Asian mothers and no Asian male toddlers is a very questionable age composition.

Use cases

The following use cases are examples of how decision makers use decennial data to inform their decisions.

Use case	Population change between 2000 and 2010 in places
Measures	Change in population between 2000 and 2010
Geography	Places (SUMLEV 160)
Universe	All incorporated places in New York State (N=616)
Tables	P1 (Total Population) and estimates base from Population estimates 2000-2010 intercensal estimates
Motivation	The decennial census is important for city and village governments to examine change in their community. The Census Bureau itself doesn't recommend using the ACS for counts and population estimates only provide total population counts and none of the characteristics.
Conclusion	Many places in New York would have drawn very different conclusions about population change in their village/city. For many people that are familiar with local change the demonstration data would have led to confusion and a lack of trust in the Decennial Census as the counts clearly don't come close to the expectations.

Analysis



Not in the chart:

Name	Pop 2000	2010 SF1	SF1 change	2010 DHC	DHC change	Perc point difference
Ames village	173	145	-16.2%	98	-43.4%	-27.2%
Constableville village	305	242	-20.7%	149	-51.1%	-30.5%
Dering Harbor village	13	11	-15.4%	32	146.2%	161.5%
Kiryas Joel village	13,139	20,175	53.6%	20,002	52.2%	-1.3%
New Square village	4,417	6,944	57.2%	6,941	57.1%	-0.1%
Ocean Beach village	138	79	-42.8%	82	-40.6%	2.2%
Saltaire village	43	37	-14.0%	93	116.3%	130.2%
West Hampton Dunes village	11	55	400.0%	53	381.8%	-18.2%

In 113 places (out of 616) the 2010 DHC indicated a change 5% or more below the change according to the 2010 SF1. In 34 places the 2010 DHC indicated a change 5% or more above the SF1 change.

On average the 2010 DHC change was -1.4 percentage points lower. Looking at the absolute differences, the 2010 DHC change was on average 4.3 percentage point different from the SF1 change.

In 87 places the DHC indicated a change in population in the opposite direction of the SF1. An example is Theresa village: according to the 2010 SF1 this village grew from 812 to 863 persons (+6.3%), but according to the 2010 DHC it declined from 812 to 659 (-18.8%).

Use case	Census Data Prison Adjustment
Measure	Prison population and NY-DOCSS counts
Geography	Blocks with prison population
Universe	Prison block identified by NY Legislative Task Force
Tables	P14 (SEX BY AGE FOR THE POPULATION UNDER 20 YEARS)
Motivation	The NY Legislative Task Force is tasked with adjusting the Census Count to confirm with NY State Law that Federal and State prisoners should be counted at a previous address.
Conclusion	The Task Force would have been confronted with counts that didn't match the administrative records, might have exceeded capacities and mismatched sex ratio's. Because it would have been the only data to work with, they knowingly would introduce errors.

Section 1. Section 71 of the correction law states:

> 8. (a) In each year in which the federal decennial census is taken but in which the United States bureau of the census does not implement a policy of reporting incarcerated persons at each such person's residential address prior to incarceration, the department of corrections and community supervision shall by September first of that same year deliver to the legislative task force on demographic research and reapportionment the following information for each incarcerated person subject to the jurisdiction of the department and located in this state on the date for which the decennial census reports population:
>
> (i) A unique identifier, not including the name, for each such person;
>
> (ii) The street address of the correctional facility in which such person was incarcerated at the time of such report;
>
> (iii) The residential address of such person prior to incarceration (if any); and
>
> (iv) Any additional information as the task force may specify pursuant to law.
>
> (b) The department shall provide the information specified in paragraph (a) of this subdivision in such form as the legislative task force on demographic research and reapportionment shall specify.

The adjustment methods assume that the Department of Corrections and Community Supervision (DOCCS) administrative records for Census Day closely matches the Census counts, otherwise errors will be introduced in the communities where the prisons are located.

Data from the 2010 adjustments can be downloaded through: https://www.latfor.state.ny.us/data/

In this analysis we examined the total counts and counts by sex. The Mean Absolute Percentage Difference was 6.6%, meaning that on average an error of 6.6% was introduced for the prison population. The largest numeric difference was for the Elmira Correctional Facility (SF1: 1787, Demo Data: 1668), the largest percentage difference was for Beacon Correctional Facility (Sf1: 168, Demo Data: 242).

Beacon Correctional was also an extreme case in difference in sex composition. In the SF1 data there were 167 females and 1 male counted in this female prison. The demonstration data estimates 2 females and 240 males. The DOCCS administrative records listed 167 persons, presumably all or nearly all females. If LATFOR in 2010 would have been confronted with this data:

- They would have wondered why they only had 167 records, while the count would have indicated 242.
- The method probably would have removed 242 persons (240 male, 2 female) from this block and at the most 167 persons (probably all female) added to their previous address

Conclusion: the Task Force would have been confronted with counts that didn't match the administrative records, might have exceeded capacities and mismatched sex ratio's. Because it would have been the only data to work with, they knowingly would introduce errors.

Use case How big is the next generation of elementary students in my school district compared with the current generation

Measure $Generation\ change = 100\% * \frac{\text{population age 0 through 5} - \text{population age 6 through 11}}{\text{population age 6 through 11}}$

Geography School districts in NY State (SUMLEV 950, 960, 970)

Universe School districts with at least 100 population age 6 through 11 in the SF1 data

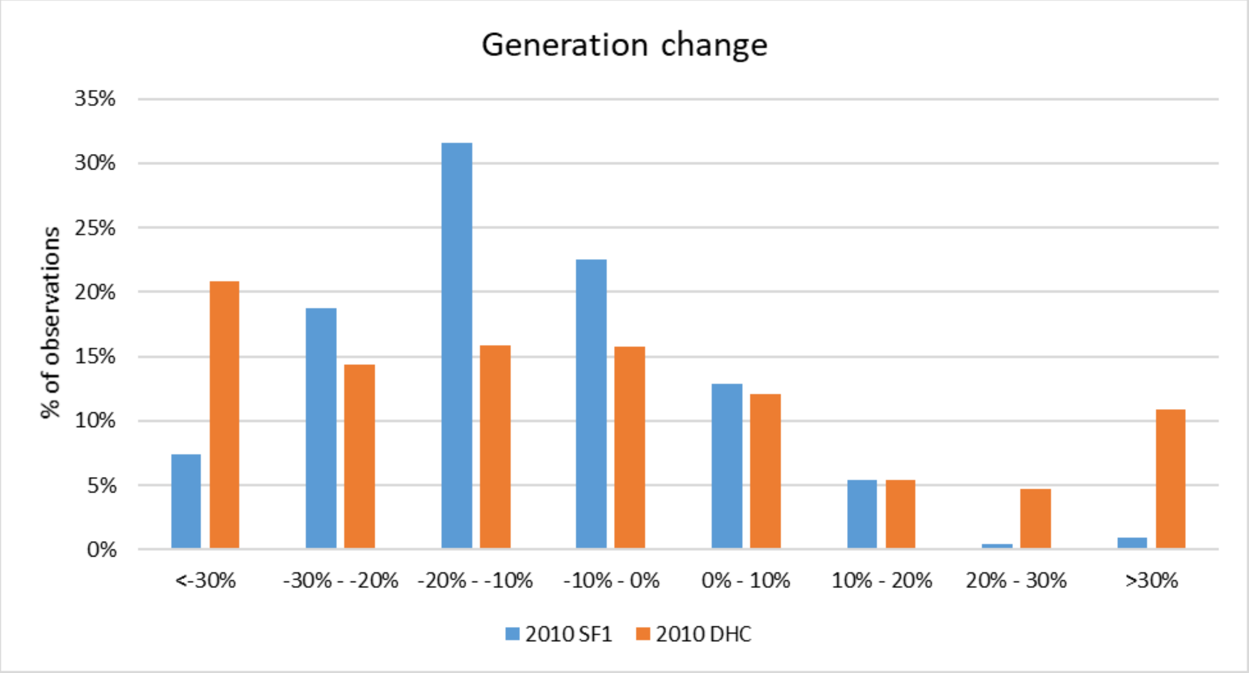
Tables P14 (SEX BY AGE FOR THE POPULATION UNDER 20 YEARS)

Motivation In 6 years time the current generation of elementary school students is replaced with the next generation. A school board want to use the census to gauge whether the next generation is smaller or larger to anticipate growth or decline.

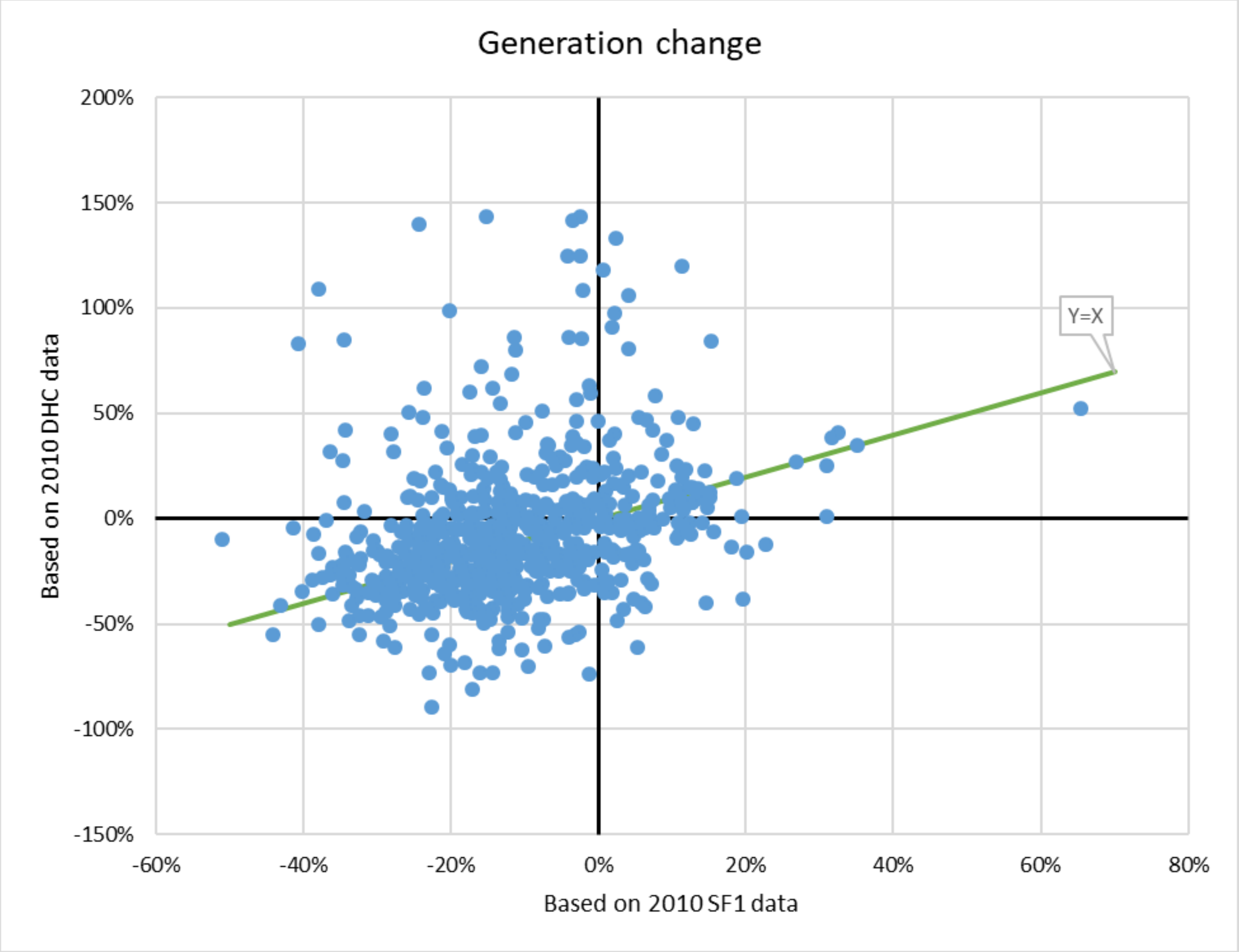
Conclusion Many School District will draw very different conclusions based on the demonstration DHC data, many will also put it aside as the data just doesn't make sense for them.

Summary Statistics and histogram

	2010 SF1	2010 DHC
Min	-51.0%	-89.5%
Max	65.5%	584.0%
Mean	-11.3%	-4.3%
StDev	0.5%	1.7%
N	661	661
Correlation	0.195	



Scatterplot



Outside the chart:

	2010 SF1 data			2010 DHC table		
School district	0-5	6-11	Generation change	0-5	6-11	Generation change
Clifton-Fine Central School District	135	158	-14.6%	171	25	584%
Downsville Central School District	118	146	-19.2%	291	63	362%
Elizabethtown-Lewis Central School District	124	146	-15.1%	277	64	333%

Jan Vink presented more results from these analyses on School Districts at the CNSTAT workshop on 2020 Census Data Products: Data Needs and Privacy Considerations[4].

[4] CNSTAT Workshop on 2020 Census Data Products: Data Needs and Privacy Considerations https://sites.nationalacademies.org/DBASSE/CNSTAT/DBASSE_196518

Use case The NY Assembly Legislative Commission on Rural Resources wants to examine difference between Urban and Rural New York

Measures Percent population in Rural Areas
Median Age
Percent population that is Non-Hispanic White Alone

Geography State and Counties (SUMLEV 040, 050), Urban and Rural geographic components (GEOCOMP 00, 01, 43)

Universe NY State and NY Counties outside the 5 New York City boroughs that have zero or very few people residing in Rural areas

Tables P5 (HISPANIC OR LATINO ORIGIN BY RACE)
P13 (MEDIAN AGE BY SEX)

Motivation New York State Legislature established the Legislative Commission on the Development of Rural Resources in 1982. From its mission Statement "The bipartisan Commission is charged with examining the impact of rural resources on the State's economy, reviewing existing laws and regulations as they relate to rural resources, and assessing the effectiveness of programs designed to promote rural viability." To better understand the population in the rural areas they want to use data from the Decennial Census and contrast the rural population with the urban population on a variety of demographic characteristics.

Conclusion There seem to be a bias in the DHC data that reduces the differences between Urban and Rural populations.

Analysis

Percent population in rural areas

	New York State		
Source	Total population	In Rural areas	Percent of total
2010 SF1	19,378,102	2,349,997	12.1%
2010 DHC	19,378,102	2,368,995	12.2%
Difference	0	18,998 (0.8%)	

Summary county comparisons	Rural population	Percent population in rural areas
Counties where DHC > SF1	53	48
Counties where DHC = SF1	0	1
Counties where DHC < SF1	4	8
Counties with missing data		

Statewide the 2010 DHC would have indicated a slightly larger rural population. The same can be said for most of the counties. Suffolk County saw the largest difference in rural population (+3.4%). The share in Allegany county is 78.7% in the SF1 data and 79.8% in the DHC data, making it the largest difference in share.

Median age

	New York State			
Source	Median age	Median age in urban areas	Median age in rural areas	Median age gap
2010 SF1	38.0	37.2	43.4	6.2
2010 DHC	38.0	37.3	42.8	5.5
Difference	0	0.1	-0.6	-0.7

Summary county comparisons	Median age	Median age in urban areas	Median age in rural areas	Median age gap
Counties where DHC > SF1	27	49	7	5
Counties where DHC = SF1	20	3	3	1
Counties where DHC < SF1	10	4	47	50
Counties with missing data		1	0	1

Statewide the 2010 DHC would have indicated a slightly older urban population and a younger rural population. The age gap between rural and urban population is 5.5 years, whereas SF1 indicated a gap of 6.2 years. Similar conclusions can be drawn for most of the counties. In three counties SF1 indicated that the rural population was older than the urban population, but DHC data would show that the urban population was older in these counties. Rockland County saw the biggest decrease in the age gap (42.4 – 36.7 = 5.7 in SF1) to (37.9 – 36.8 = 1.1 in DHC)

Percent Non-Hispanic White alone

	New York State		
Source	% NHW	%NHW in urban areas	%NHW in rural areas
2010 SF1	58.3%	53.5%	93.1%
2010 DHC	58.3%	53.6%	92.4%
Difference	0	0.04 pp	-0.62 pp

Summary county comparisons	% NHW	%NHW in urban areas	%NHW in rural areas
Counties where DHC > SF1	23	48	4
Counties where DHC = SF1	0	0	0
Counties where DHC < SF1	34	8	53
Counties with missing data		1	0

Non Hispanic White population shares are reported lower in most of the rural areas and higher in most of the urban areas. Statewide the percentage Non Hispanic White population in rural areas is 93.1% in SF1 and 92.4% in the demonstration products.

Use case Distribution of Albany County Sales Tax revenue towards cities and towns

Measures Local Revenues from County Sales Tax
Tow/City share of population

Geography Counties and MCD (SUMLEV 050, 060)

Universe Cities and towns in Albany County

Tables P1 (Total Population)

Motivation The Sales Tax rate in New York consist out of two components; a state tax rate and a county tax rate. The county tax rate varies throughout the state. How the revenue from the county sales tax is divided between county and sub county entities also differs between county; In some counties it is all for the county, in other counties shares that go to the towns/cities depend on property values and in four counties the share of the county sales tax that goes to towns and cities depends on the latest Decennial Census. The sharing agreement in Albany County for example states "The County retains 60% and distributes 40% to the cities and towns on the basis of published decennial census population figures."

Conclusion The distribution of sales tax in Albany County would have been influenced because of the DAS. The table represent the differences for a single year, but since shares are calculated solely on the Decennial Census, these differences will occur every year in the decade. As percent of the total budget these differences seem small, but the lost revenues in for example Colonie town can have a real impact.
In other counties, MCD share of the total county population can differ more than within Albany. There are 8 MCD where the difference in share between SF1 and DHC was more than 0.5 percentage points.

Analysis

Albany County collects annually around 275 million dollar in sales tax. 40% of that 110 million is distributed towards towns and cities according to the local sharing agreement. For Albany MCD's this is around 25% of their total revenue.

MCD	SF1 pop	DHC pop	SF1 Share	DHC share	Share of sales tax (SF1)	Share of sales tax (DHC)	Difference
Albany city	97,856	97,941	32.2%	32.2%	35,384,676	35,409,359	24,683
Berne town	2,794	2,795	0.9%	0.9%	1,010,309	1,010,498	189
Bethlehem town	33,656	33,612	11.1%	11.0%	12,169,991	12,152,004	-17,988
Coeymans town	7,418	7,458	2.4%	2.5%	2,682,345	2,696,348	14,003
Cohoes city	16,168	16,316	5.3%	5.4%	5,846,340	5,898,848	52,508
Colonie town	81,591	81,422	26.8%	26.8%	29,503,261	29,437,119	-66,142
Green Island town	2,620	2,648	0.9%	0.9%	947,391	957,352	9,961
Guilderland town	35,303	35,220	11.6%	11.6%	12,765,545	12,733,356	-32,189
Knox town	2,692	2,678	0.9%	0.9%	973,426	968,198	-5,228

New Scotland town	8,648	8,663	2.8%	2.8%	3,127,112	3,132,001	4,889
Rensselaerville town	1,843	1,858	0.6%	0.6%	666,428	671,737	5,309
Watervliet city	10,254	10,275	3.4%	3.4%	3,707,841	3,714,799	6,959
Westerlo town	3,361	3,370	1.1%	1.1%	1,215,336	1,218,382	3,046
Total	304,204	304,256	100.0%	100.0%	110,000,000	110,000,000	0

Use case Tract level demographic analyses

Measure Average household size, median age, child woman ratio, share Hispanic population, Occupancy rate

Geography Tracts

Universe Tracts in NY for which measure could be calculated

Tables P17, P13, P12, P12H, H3

Motivation Tracts have become a very important statistical level of geography to analyze the population, especially when it comes to analyzing spatial differences. These kinds of analyses are supporting decisions by local governments, businesses and also help researchers understand aspects of the population.

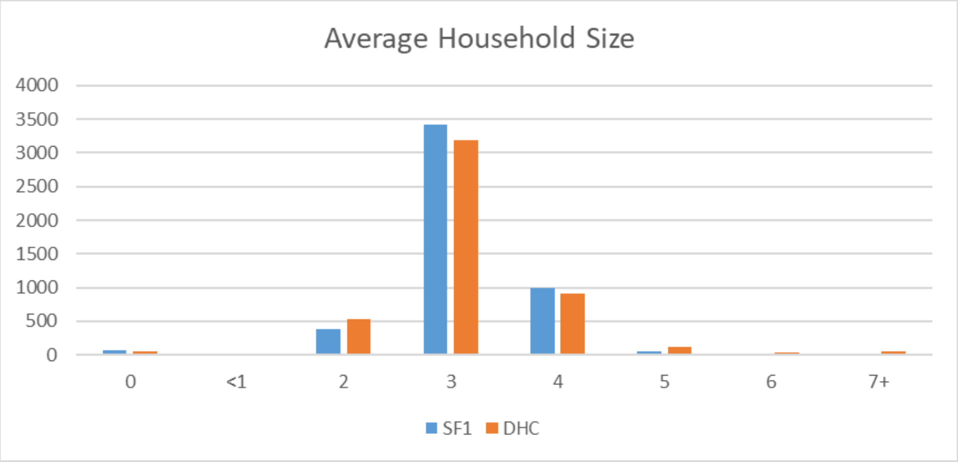
Conclusion Some of the tract level statistics are very different between the SF1 and the demonstration data. If we define high quality data as accurately describing the real world phenomena, than we need to wonder if the tract-level demonstration data can be called “high quality”.

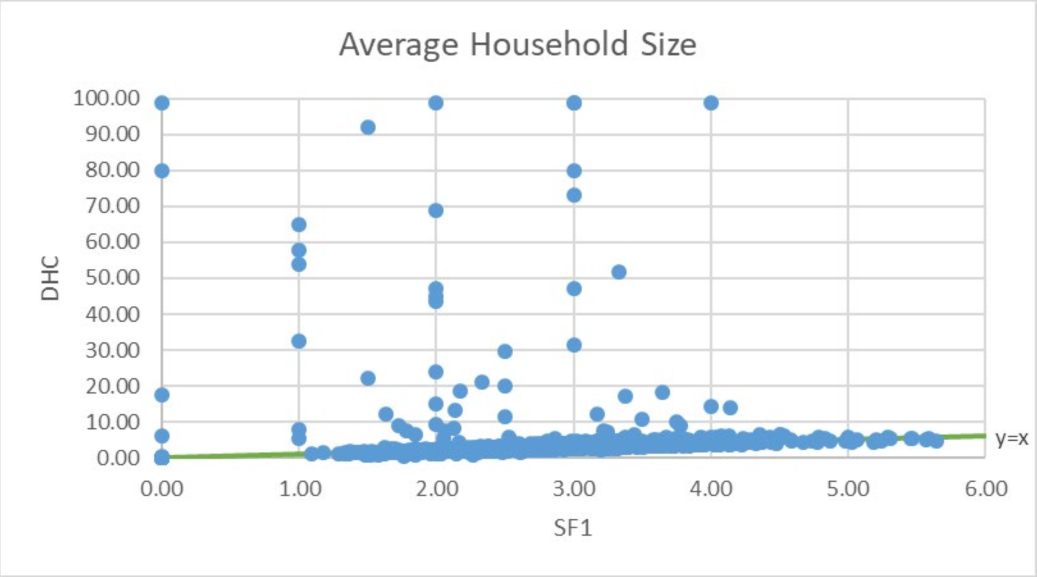
In a the final federal register notice on the 2020 census tract criteria, the Census Bureau mentions: “the primary purpose of census tracts is to help provide high-quality statistical data about the population”. In this use case we compare tract level data from SF1 with tract level data from the demonstration data.

Results:

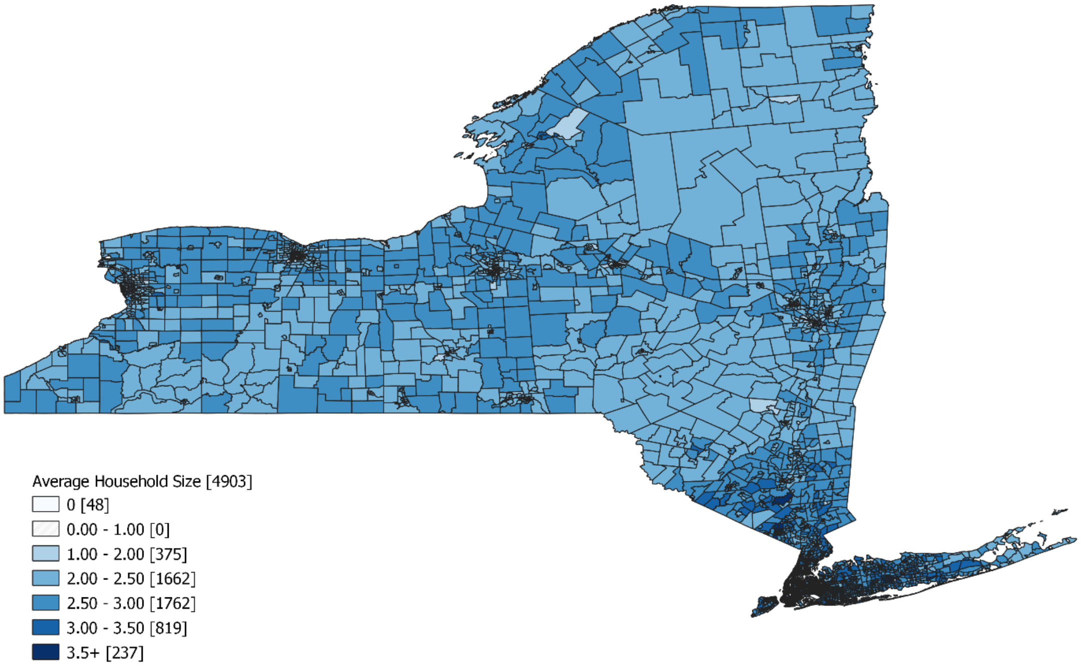
Tract level demographic analyses: Average household size

SF1		*DHC*	
Mean	2.645854	Mean	2.979782
Median	2.6	Median	2.581353
Standard Deviation	0.513876	Standard Deviation	4.595442
Range	4.64	Range	98.64924
Minimum	1	Minimum	0.350763
Maximum	5.64	Maximum	99
Count	4855	Count	4860

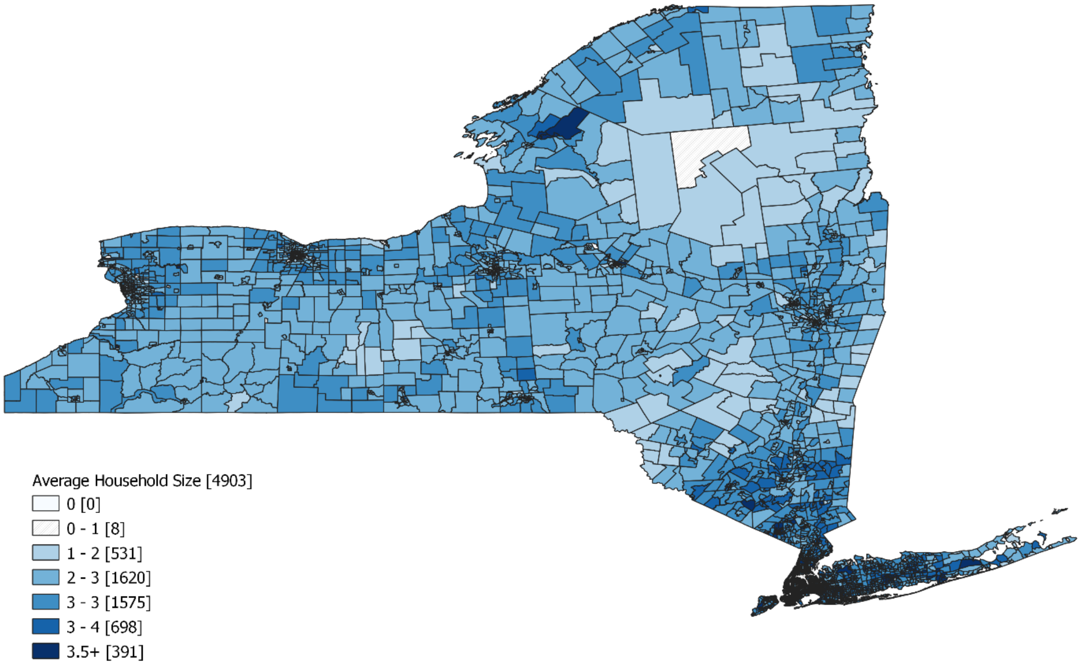

Average Household Size
4000
3500
3000
2500
2000
1500
1000
500
0
0
<1
2
3
4
5
6
7+
■ SF1 ■ DHC


Average Household Size
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
DHC
0.00
1.00
2.00
3.00
4.00
5.00
6.00
SF1
y=x

Average Household Size by Census Tract SF1 2010



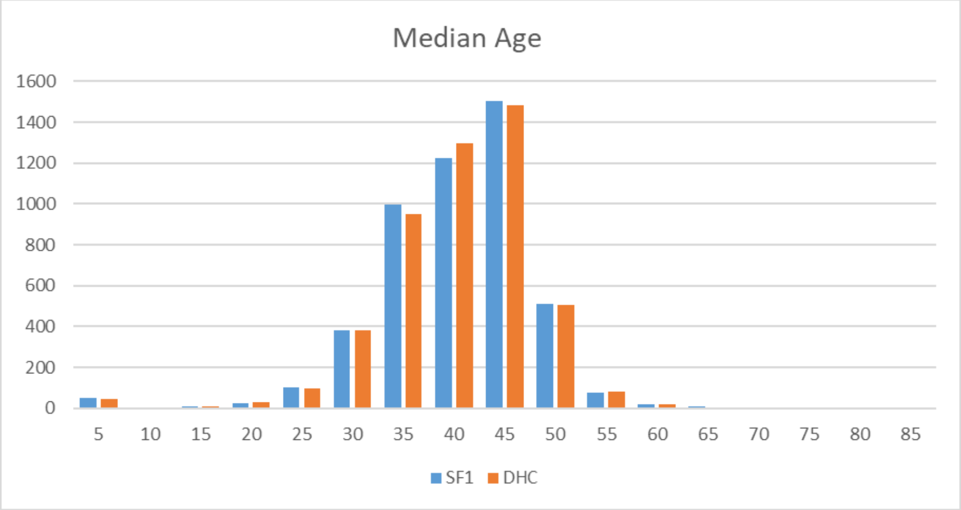
Average Household Size by Census Tract DHC 2010

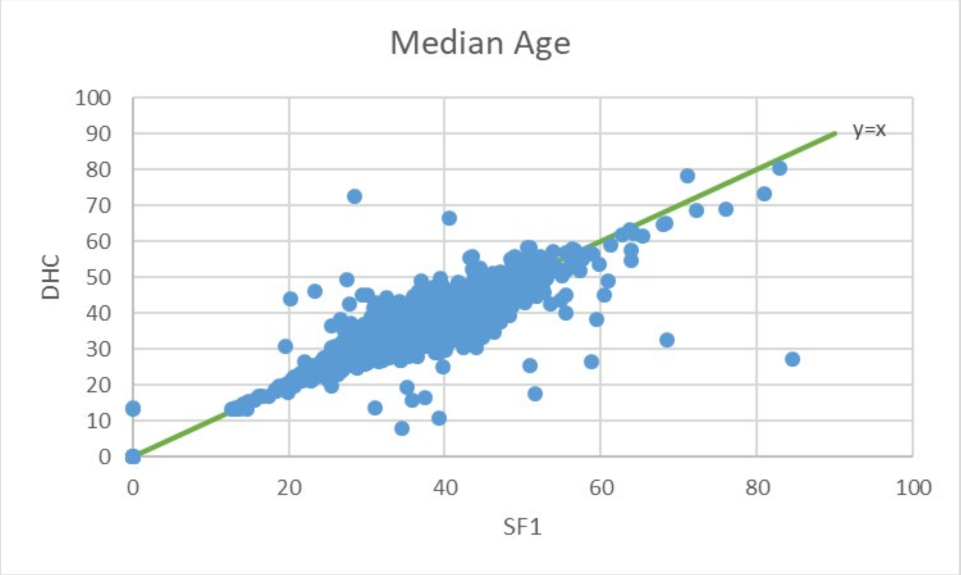


Tract level demographic analyses: Median Age

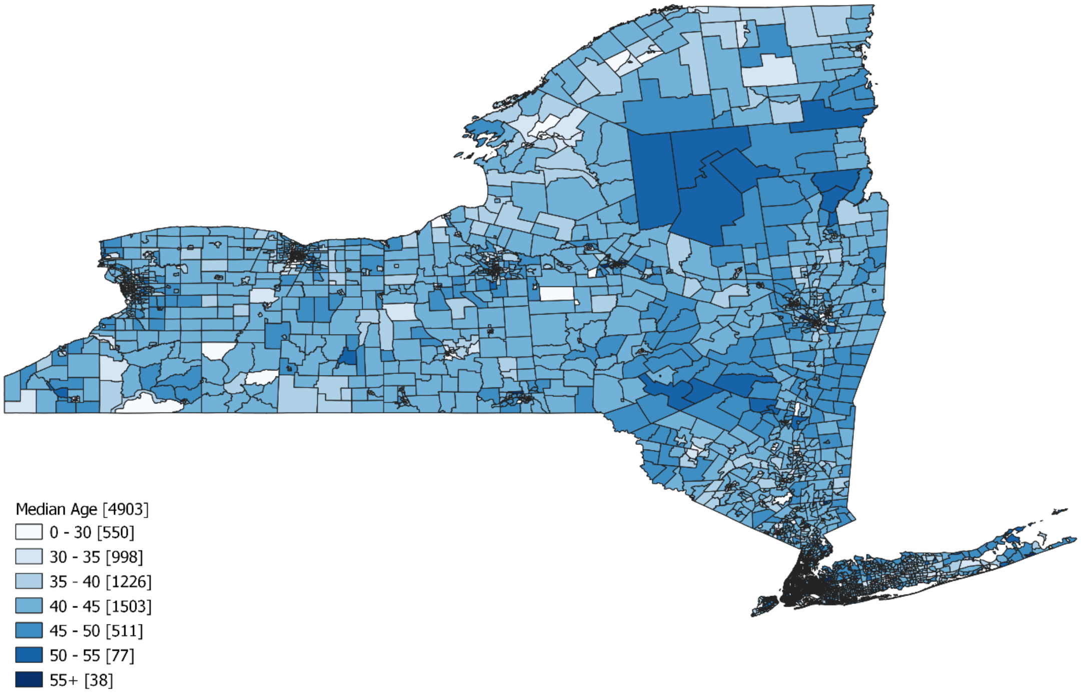
Median Age

SF1		*DHC*	
Mean	38.3169	Mean	38.25784
Median	39	Median	38.9
Standard Deviation	6.728707	Standard Deviation	6.697923
Range	71.8	Range	72.6
Minimum	12.7	Minimum	7.8
Maximum	84.5	Maximum	80.4
Count	4870	Count	4872

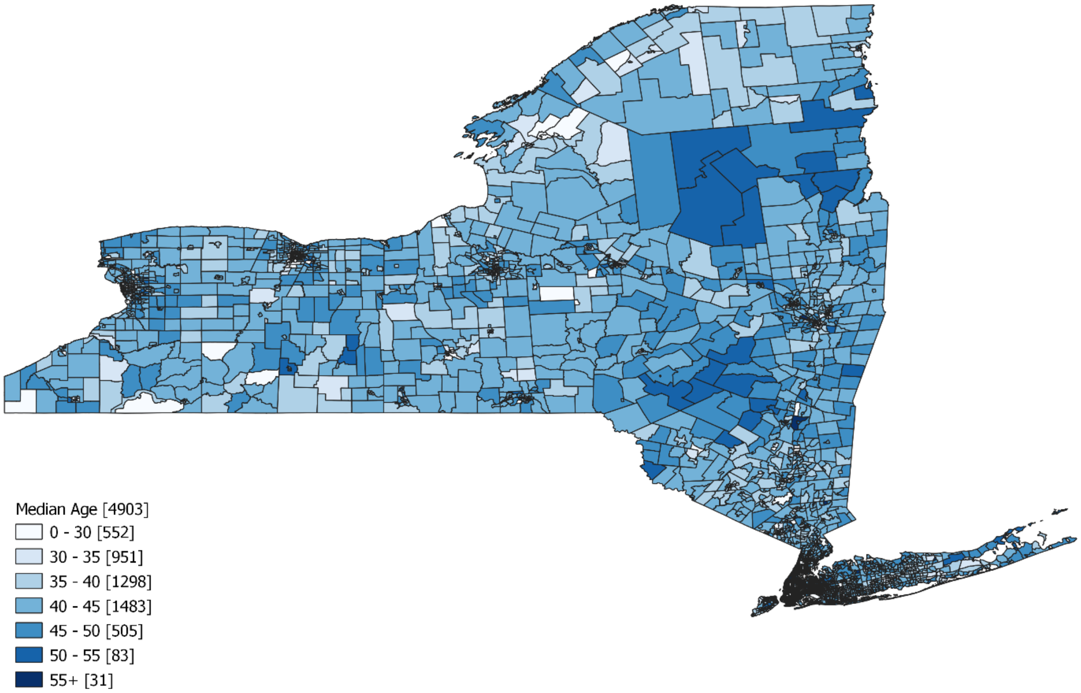




Median Age by Census Tract SF1 2010



Median Age by Census Tract DHC 2010

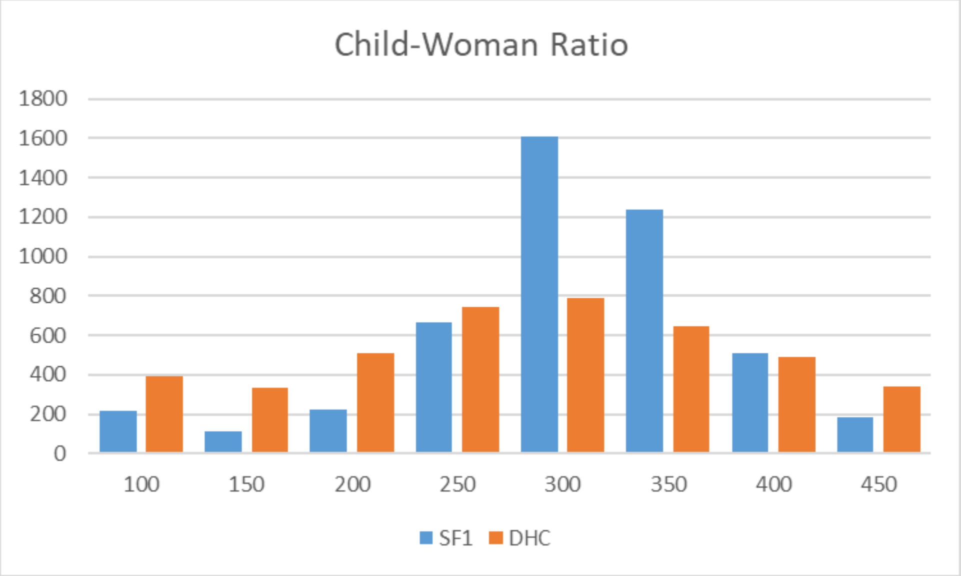


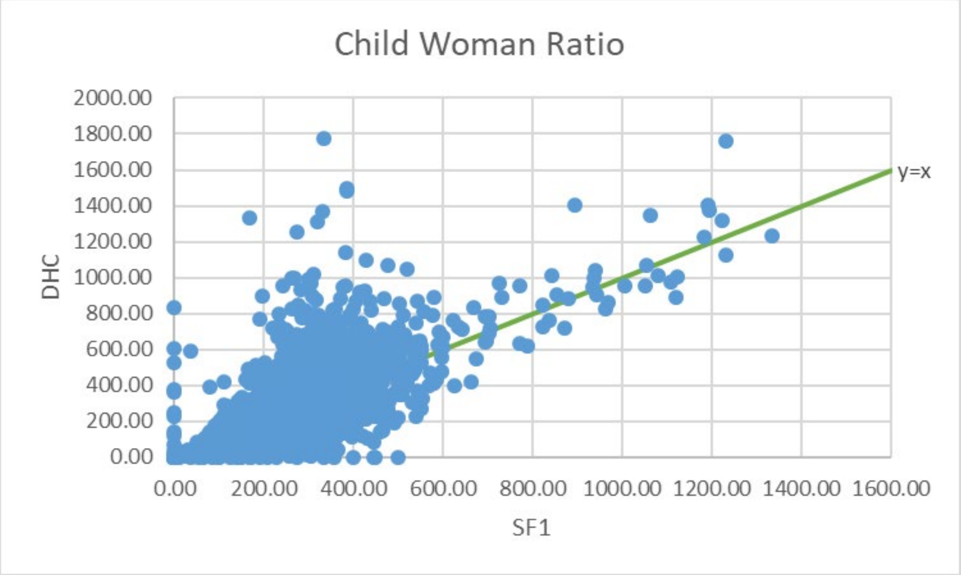
Tract level demographic analyses: child-woman ratio

Measure: $child - woman\ ratio\ = \frac{\text{children under 5}}{\text{women 15−44}} * 1000$

Child-Woman Ratio

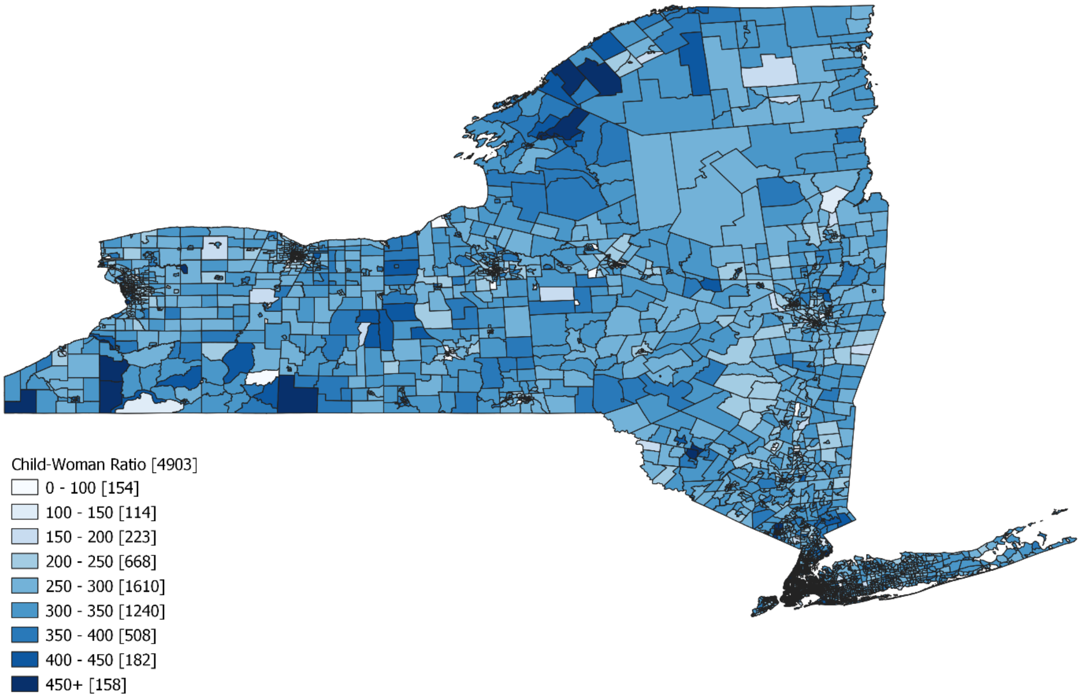
SF1		*DHC*	
Mean	295.353	Mean	309.2855
Median	291.0798	Median	283.4747
Standard Deviation	99.37869	Standard Deviation	228.077
Range	1333.208	Range	10999.31
Minimum	0.727008	Minimum	0.691085
Maximum	1333.935	Maximum	11000
Count	4821	Count	4808



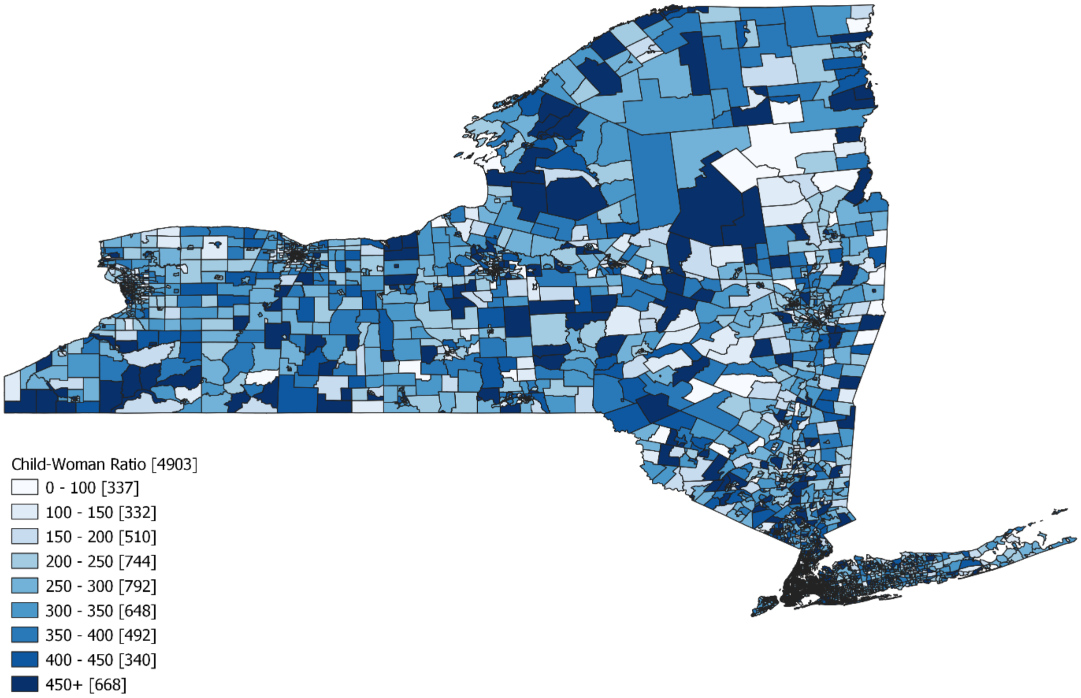


The above figure excludes two outliers where the difference between the SF1 2010 and DHC 2010 census tracts was 2,466 and 10,833.

Child-Woman Ratio by Census Tract SF1 2010

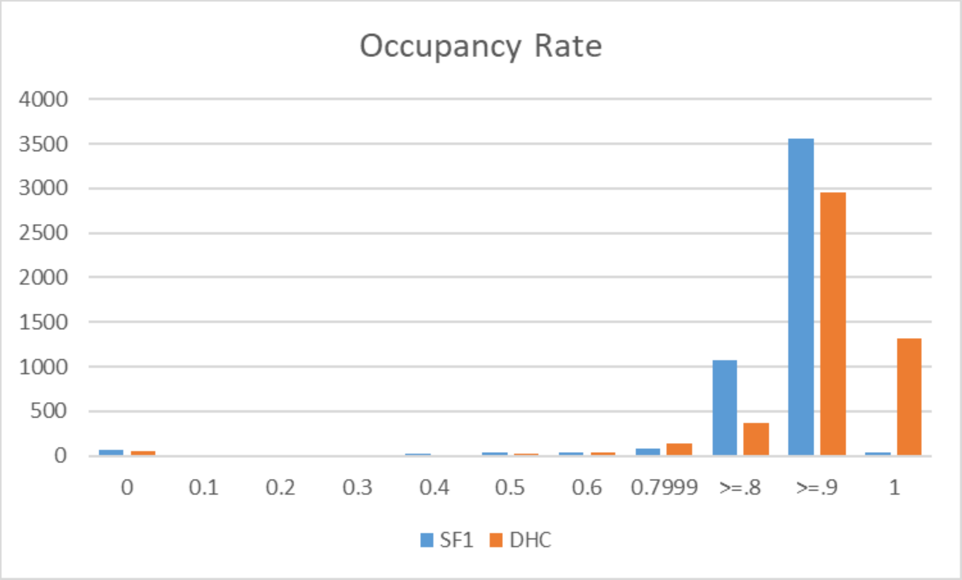


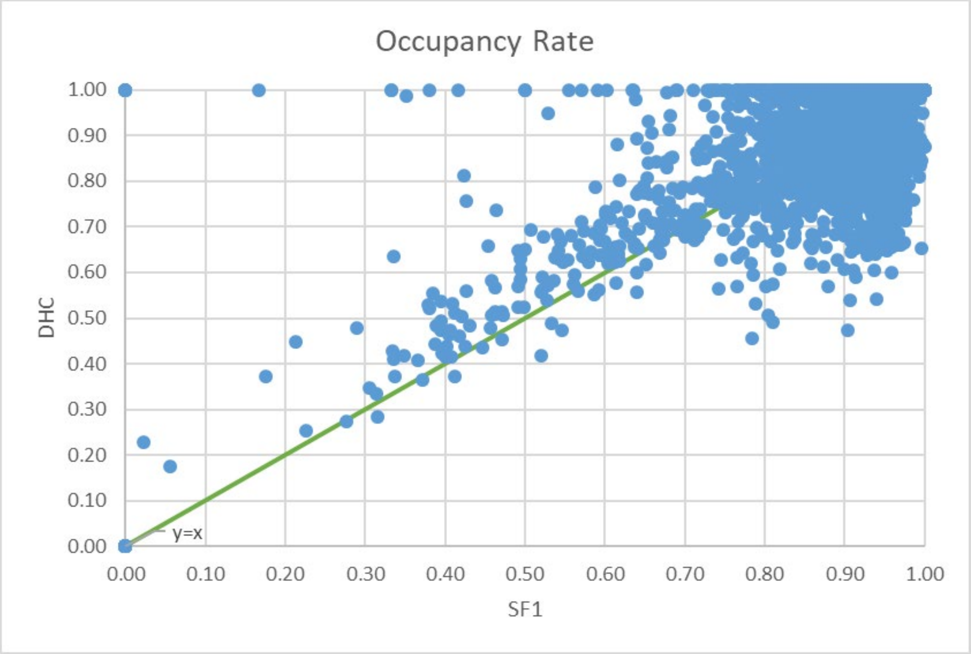
Child-Woman Ratio by Census Tract DHC 2010



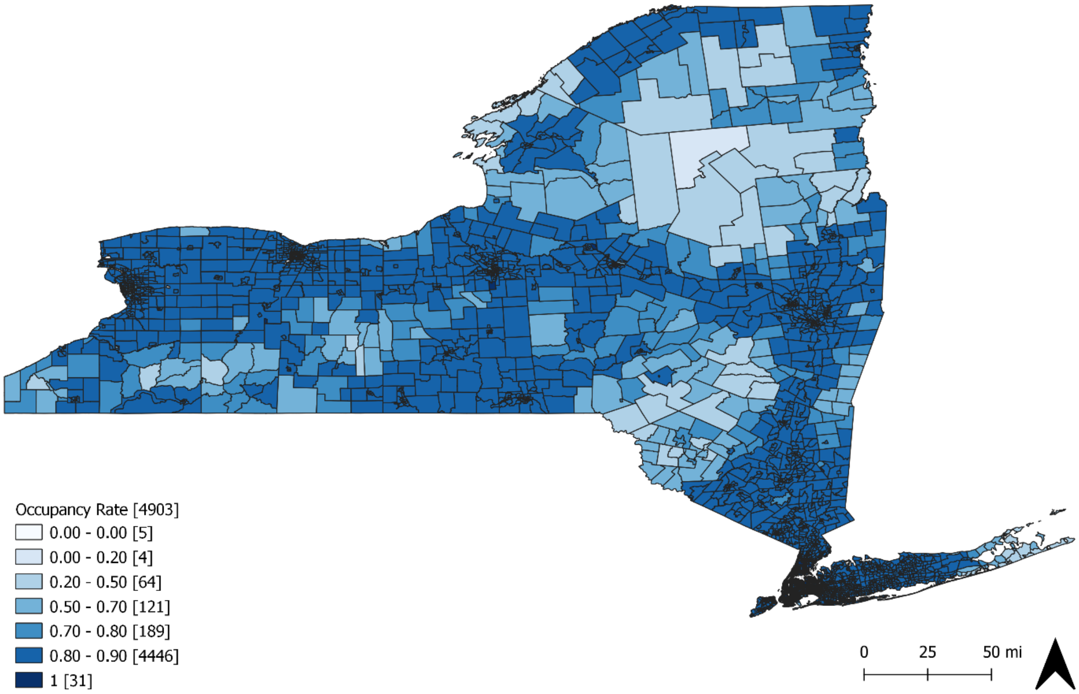
Tract level demographic analyses: Occupancy Rate

SF1		*DHC*	
Mean	0.895569	Mean	0.910136
Median	0.934368	Median	0.961721
Standard Deviation	0.13884	Standard Deviation	0.145019
Range	1	Range	1
Minimum	0	Minimum	0
Maximum	1	Maximum	1
Count	4919	Count	4919

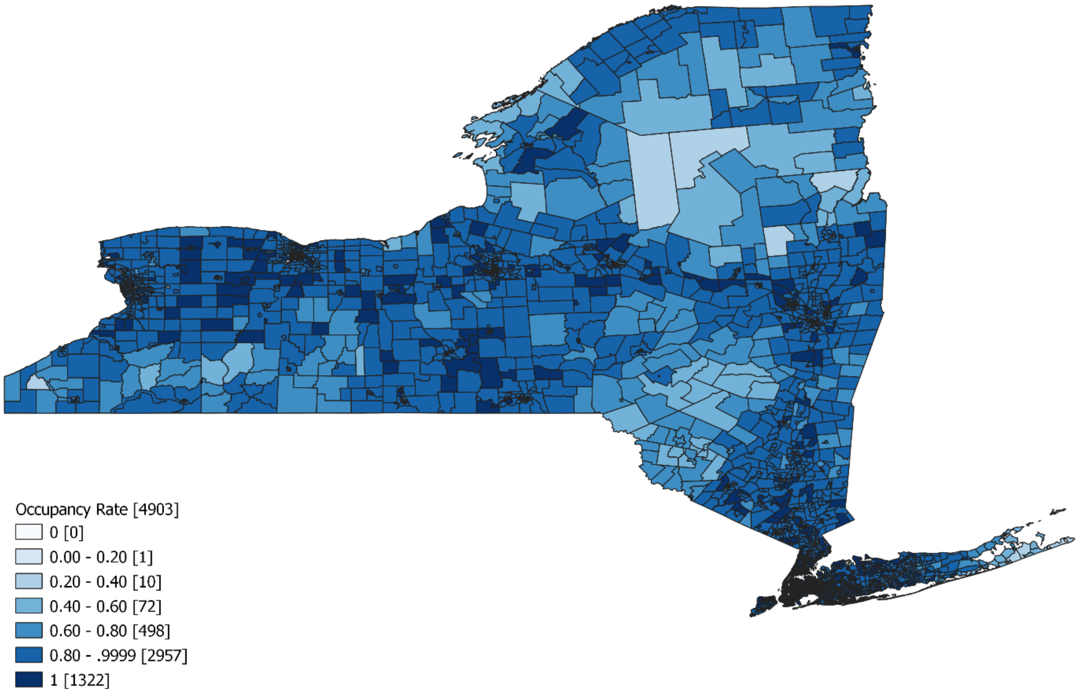




Occupancy Rate by Census Tract SF1 2010

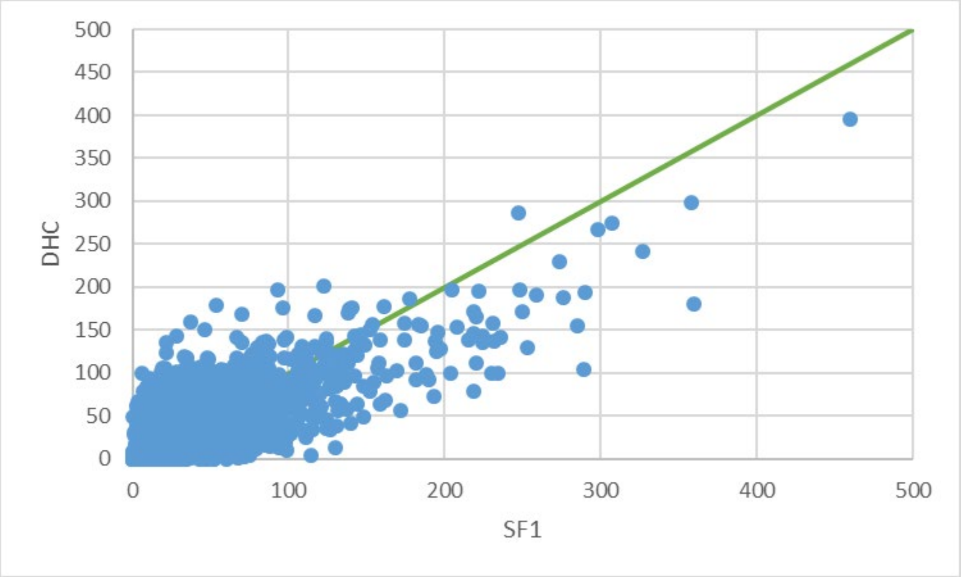


Occupancy Rate by Census Tract DHC 2010



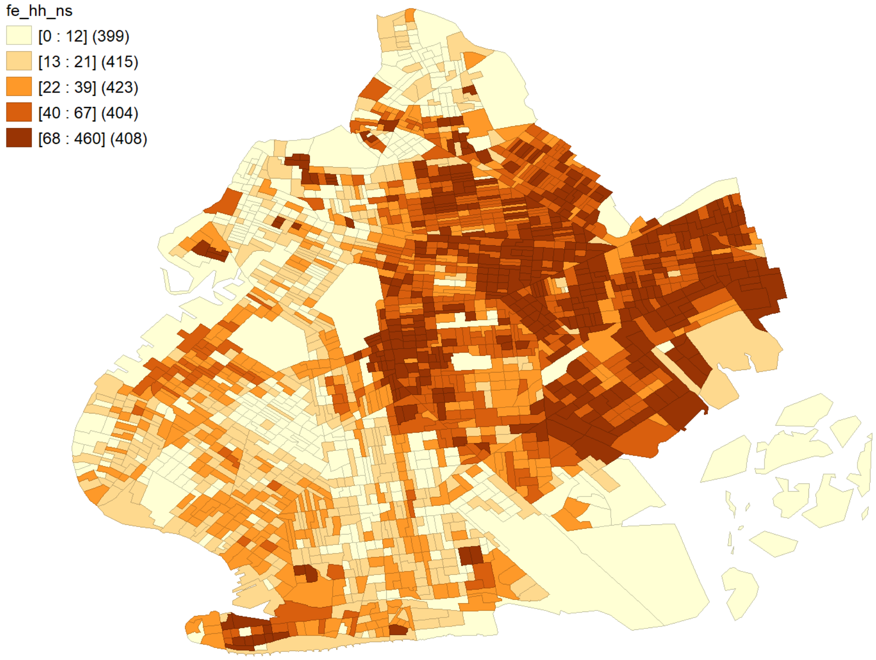
Use case Block Group level spatial analyses on female headed families with children and no spouse present

Measure	Number of family households: Female householder, no husband present: With own children under 18 years
Geography	Block groups
Universe	Block groups in Kings County [Brooklyn]
Tables	P19
Motivation	Family households with a female householder, no spouse present and with children are most vulnerable and have high poverty rates. To deliver services to this group spatial analyses are done to find neighborhoods with large counts of such families.
Conclusion	Whereas in SF1 block groups with large numbers of female headed families are clustered, in the demonstration data block groups with large numbers are much more scattered throughout the borough.

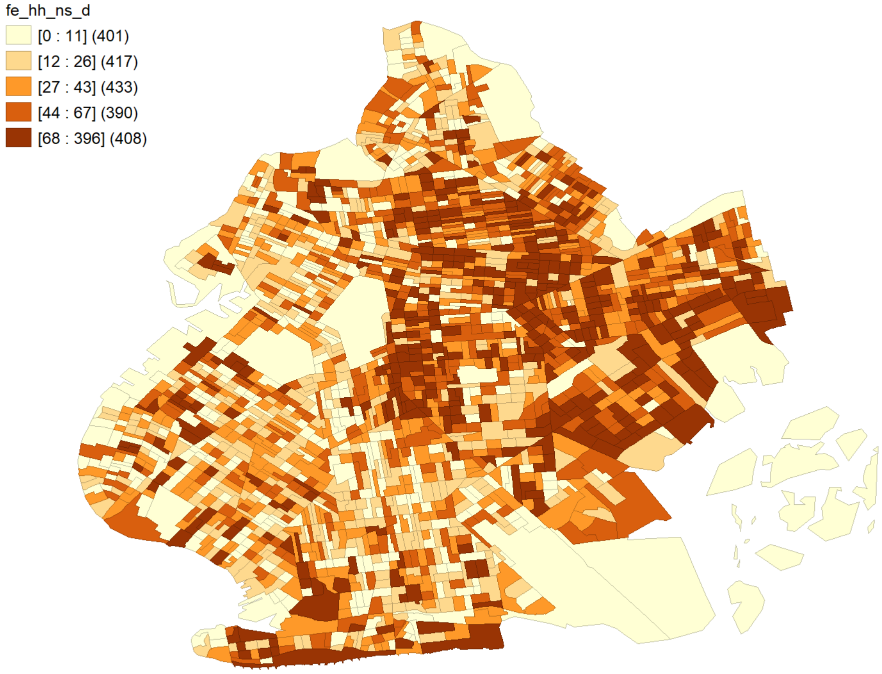


We calculated the Moran's I for both datasets (Distance matrix based on Queen contiguity-order 1)

In the SF1 the Moran's I was 0.484, in the demonstration data DHC it was 0.263, an indication that a lot of the spatial autocorrelation got lost.



Number of female headed families with children and no spouse present (Source: SF1)



Number of female headed families with children and no spouse present (Source: DHC)

Use case Characteristics of the hard-to-count populations

The 2010 SF1 data was combined with response rates to learn about populations that might be harder to count. The Census operation in 2020 benefit from these analyses as outreach can be more targeted, which makes the total operation cheaper and more efficient. If the hard-to-count analyses would have been based on the demonstration data, much weaker relations would be found and targeting of the next Census would have been much less efficient.