Feedback on the demonstration data sets released in August 2022

As sent to the U.S. Census Bureau on September 26th 2022

Program on Applied Demographics

Cornell University

Program on Applied Demographics

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

INTRODUCTION

In 2019 the Census Bureau requested detailed input on what tables from SF1 were important to the data users, and what levels of geography were needed. Data users were asked to describe the use case, and if possible the accuracy required. Later that year the Census Bureau released the first 2010 based demonstration dataset that was produced with Differential Privacy. Before that there had been proof of concepts with applications based on the 1940 Census and with the 2018 End-to-end test in Providence, RI.

Following the 2019 demonstration data release , CNSTAT organized a workshop where a select set of papers were presented that highlighted problems with accuracy and usability if the Census Bureau had released this dataset instead of SF1.

During 2020 and 2021 the Census Bureau released subsequent demonstration datasets, attempting to address some of the short-comings that were pointed out in the initial and following rounds of feedback. At one point the Census Bureau decided to focus on the production of the PL94-171 redistricting files and this limited the fields included in the demonstration data. In August 2021 the Census Bureau released the 2020 PL94-171 data product, followed by a 2010 based demonstration dataset using the same settings.

In March/April 2022 and August 2022 the U.S. Census Bureau released new demonstration data with a complete set of tables. The August release included many more tables reflecting the feedback from data users. Between these two demonstration dataset releases was another CNSTAT workshop with selected papers highlighting issues of accuracy, privacy and usability.

The Cornell Program on Applied Demographics has been very active in responding to all feedback opportunities with the goal to get the most usable data for 2020 given the privacy parameters set by the US Census Bureau. We participated in both CNSTAT workshops and have had to develop new metrics as Mean Averages often don't tell the whole story.

Attached is the feedback document we sent to the Census Bureau in reaction to the August demonstration data release, which is intended to be the last one before final decisions are made.

Data users were given 30 days to respond, and the feedback's only intended audience is the U.S. Census Bureau. The resulting document contains only brief notes/discussion and assumes deep knowledge of the Top Down Algorithm.

Feedback on the demonstration data sets released in August 2022

Authors: Jan Vink and Leslie Reynolds
Cornell Program on Applied Demographics
September 26, 2022
Email: padinfo@cornell.edu

1 INTRODUCTION AND CONCLUSIONS

On August 25'th the Census Bureau released a new set of demonstration data with tables planned for the DHC. With only 30 days available for feedback our analyses cover a lot of topics, but due to the time constraint not all analyses could go as deep as we would help liked and although every effort is taken to present our results clearly, there was no time for finishing touches.

We didn't attempt to compare this release with the previous release as we wanted to focus our comments on accuracy and usability of the August release.

General impressions and conclusions:

- There are some big differences in accuracy and usability between incorporated places and unincorporated places of the same size. This is very worrisome as many data uses will require custom built geographies. CDPs (unincorporated places) are examples of custom built geographies with a Census Program (PSAP) supporting this custom build. Often these custom built geographies are relatively small, less than a few thousand people and the CDP analysis show that there often big differences between the demonstration data and SF1 in those geographies.
- The Census Bureau already recognized some shortcomings in the sex/age distributions, but besides those, the age/sex distributions for the total population look usable for on/near-spine geographies (blocks groups and above) above about 1,000 persons.
- Differences in median age get smaller with growing population size.
- Differences in median age for off-spine geographies (like CDPs) are much bigger than for the on/near-spine geographies.
- Many geographies have an multi-age age range for which the differences between demonstration data and SF1 have the same sign, resulting in a significant aggregation of those differences impacting the usability.
- Breaking the link between persons and households generates many inconsistencies in the data. Census data users use various indicators that rely on the data of both households and persons:
    - Persons per Household (household population divided by number of households),
    - headship rates (householders of a certain age divided by household population of that age) are used to link age structures of the population with number of households,
    - minority home ownership (householders of a minority group divided by population of that group)

  Inconsistencies in the data require data users to derive alternative estimates that are feasible, but not longer solely based on area specific Census Data. These inconsistencies can also be seen as symptoms of distortions of the underlying distributions.

  Inconsistencies happen on all levels of geography analyzed (county and below)
- Table cells for household type and household size that generally have lower counts (less common household types and household sizes) often have very large percentage errors (e.g. over 30% of tracts have more than 10% error) which severely limits the usability of these tables.
- The householder race/ethnicity tables were mentioned in Census Bureau presentations as tables that showed lower accuracy. Our analyses confirm that and would urge the Census Bureau to improve accuracy. This problem is even more prominent under the householders of households with 3 or more generations.
- Table cells with the largest value have a negative bias causing tables to be more 'diverse' in the demonstration data.
- Various selection biases that are not well-understood exist within these data. For example: areas with mostly rental occupied houses have a positive bias for households with children, whereas areas with mostly owner occupied houses have a negative bias for households with children.

Analyses are based on the data downloaded directly from the Census Bureau servers or on data downloaded from NHGIS-IPUMS[1]. Several of the analyses are limited to New York State.

[1] David Van Riper, Tracy Kugler, and Jonathan Schroeder. IPUMS NHGIS, Privacy-Protected 2010 Census Demonstration Data, version 20220825 [Database]. Minneapolis, MN: IPUMS. 2022.

2 CROSSWALK

We appreciate the added tables, especially bringing the level of geography down for some often-used tables.

We would like to make two additional comments:

- Table PCT1 and H7 contain the same information and it might be confusing that table PCT1 is not available for all geographies, but table H7 is. I suggest eliminating table PCT1
- PCT13 (Age of population in Households) and its iterations are very useful, but is currently limited to iterations A-I. It would be even more useful if iterations J-O would be added (just like table P12)

3 AGE DISTRIBUTIONS

3.1 RESEARCH QUESTION:

How do differences in median age vary by type of geography, size of population, sex and race/ethnicity?

3.2 CONCLUSIONS:

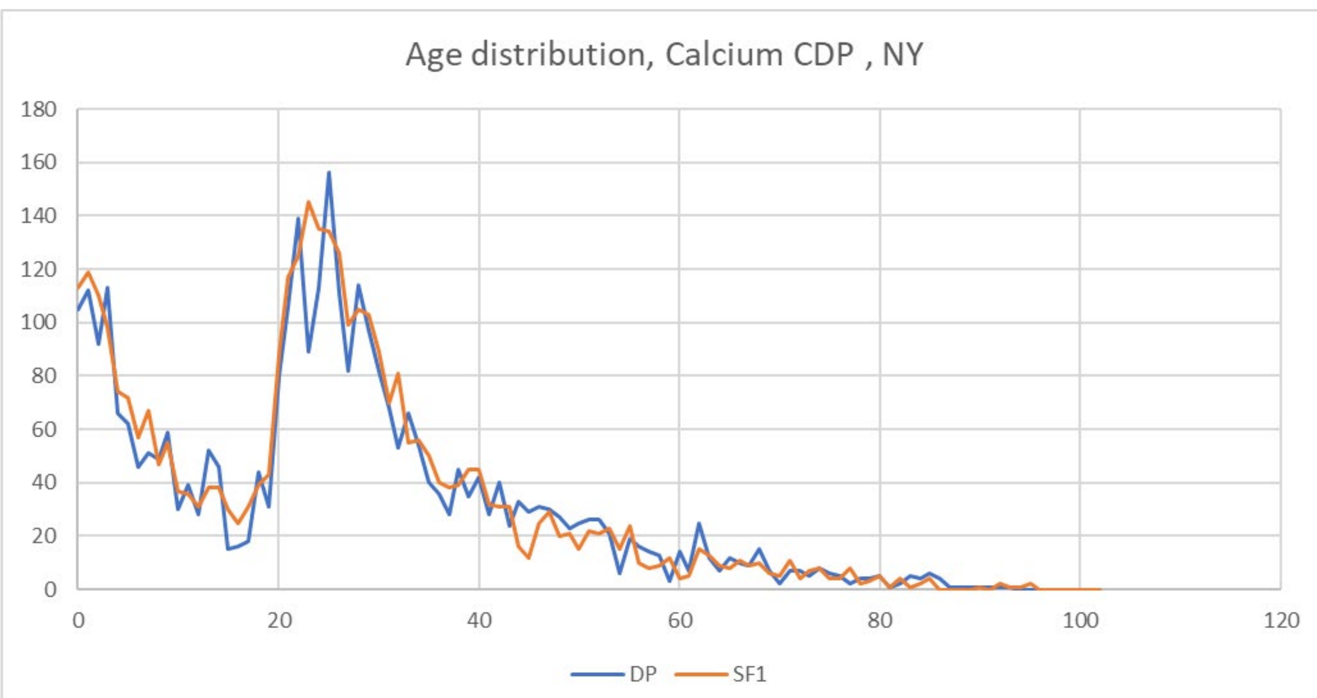
- Very small population groups (< 200) seem to have a significant negative bias
- Unincorporated places (CDPs) have larger differences compared to Incorporated places (villages and cities).
- CDPs below 2,000 population often have absolute difference of 2 year or more. This is rare among other geographies of size 500 or more
- The difference between CDPs and villages is most prominent under NH White Alone populations, not as much difference for Black Alone or Hispanic populations
- For population groups larger than 1,000 the mean absolute difference is mostly less than 1 year
- The mean absolute difference in median age for males or females where there are 1,000 to 2,000 males or females is smaller then total populations of 1,000 to 2,000
- The mean absolute difference for groups of two or more races that are 1,000 to 2,000 in size are larger than for other race categories that size

3.3 METHODS & METRICS

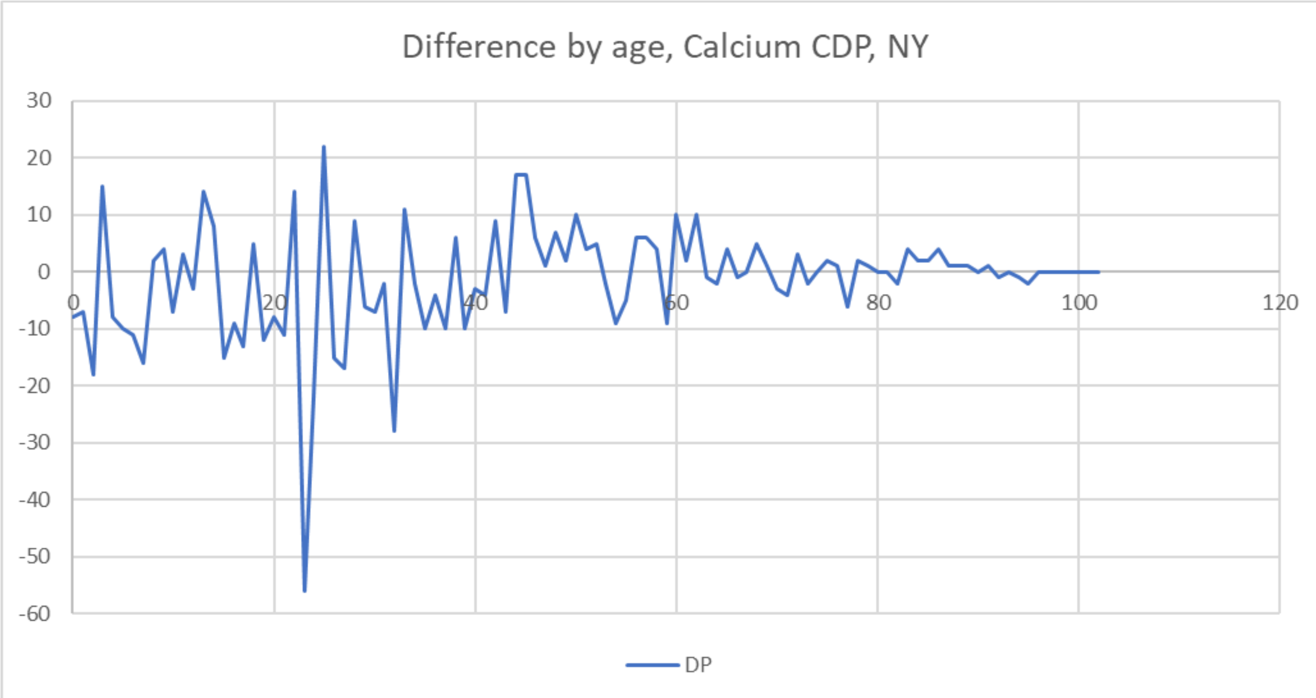
There are several ways to compare age distributions and say something about the differences.

I first use an example to explain the two methods and metrics used in this feedback. More details on the metrics are in the sections.

Example Calcium CDP:



Visually it is hard to tell if there any problems caused by DP.

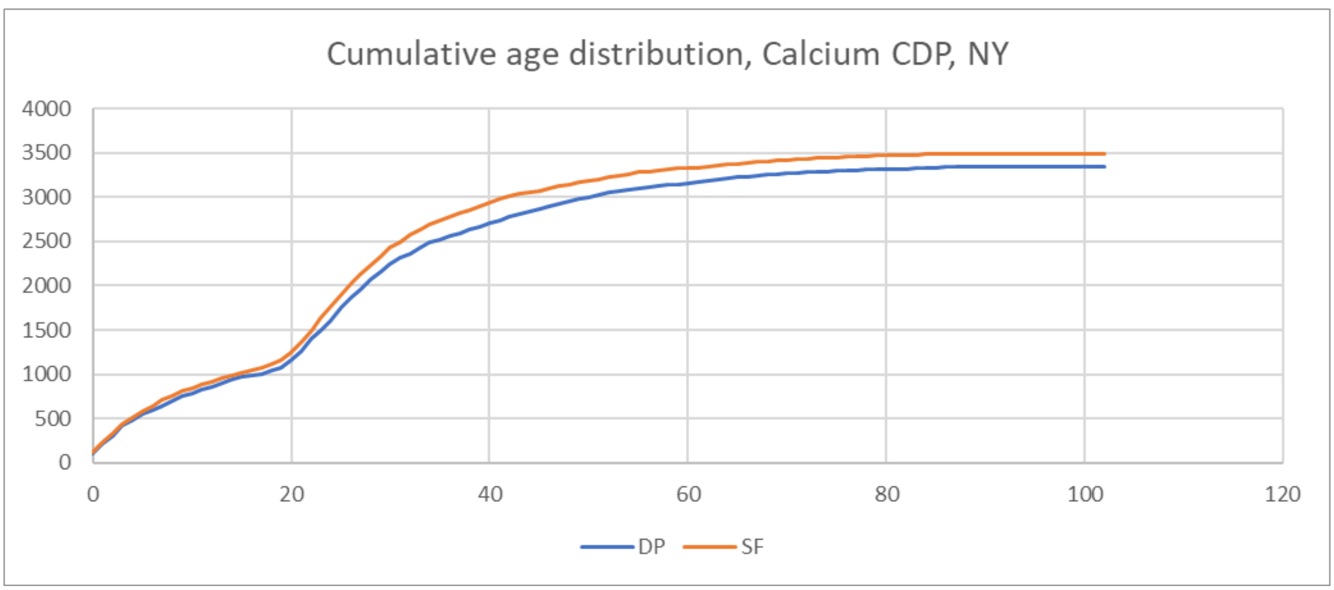


Even if we look at the differences it is hard to tell if there are any problems.

The two hypothesis I want to test are:

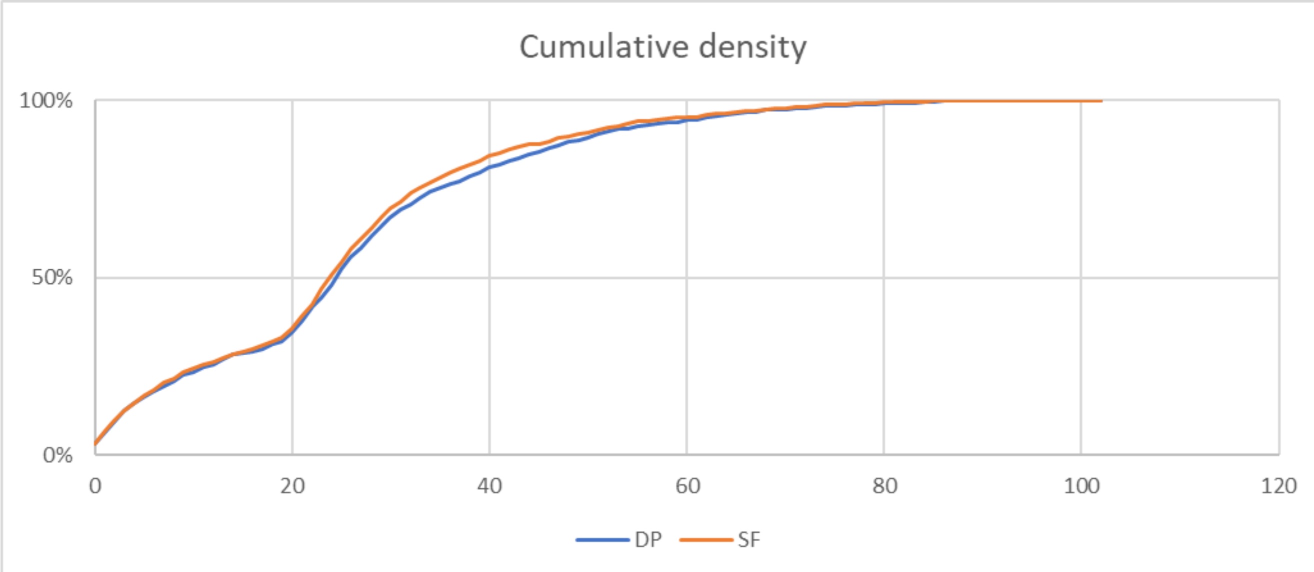
- Is there much difference in the median age?
- Is there is a group of ages for which the demonstration data yield significant different results as the published SF1 data?

To explain the methods and how they relate I start with the cumulative age distribution



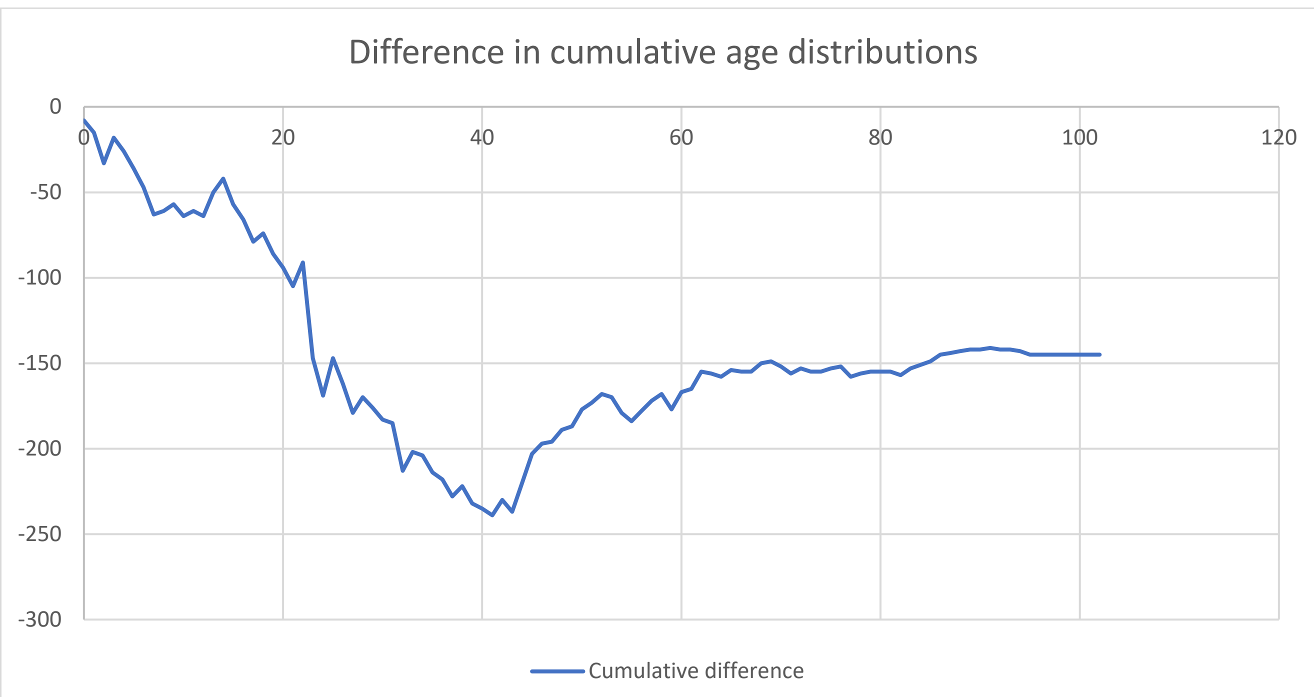
From the chart above we learn that for example in the demonstration data there were 2710 people age 0-40 and 2945 in SF1.

We can also create a cumulative density plot by dividing both distribution by the total:



From this plot we learn that the 50% of the population in Calcium is 24.5 yrs or younger in the demonstration data and 23.8 years or younger in SF1. These are the median ages and we look at the differences in median age first.

Going back to the first cumulative plot, we can also plot the difference between those two lines:



We would get the same chart if you calculate a cumulative of the errors by age.

Please note that the slope of a part of this line is mean error over that part. If we look for example between age 40 and age 60, we see a positive slope. The demonstration data counts 448 persons age 41-60 and SF1 counted 381. Indeed the positive slope corresponds with demonstration count higher then in SF1. This chart also makes it much clearer that the negative differences in the age range 0-40 far outweigh the positive differences, leading to the mentioned gap between 2710 (DP) and 2945 (SF1). Where the total population in this CDP was 4% lower in the demonstration data, for this age range the difference was 8%. In the second part of this chapter, we will be looking more into these differences.

3.4 MEDIAN AGE

3.4.1 Methodology

I only looked at geographies in New York State.

Types of geographies based on SUMLEV and first character of COUSUBCC and PLACECC considered are:

- Counties
- Incorporated places
- Unincorporated places,
- Towns & Cities (MCDs),
- Unified School Districts (SD).

Please note that cities in New York are an incorporated places (SUMLEV = 160) as well as an Subcounty (SUMLEV = 060). For these analyses they are their own category, but also included in the incorporated places.

For each of the subgroup I looked at the population size in SF1 and coded that in the following population size bins:

- 0-9 (Excluded from further analyses)
- 10-199
- 200-499
- 500-999
- 1,000-1,999
- 2,000-4,999
- 5,000-9,999
- 10,000+

3.4.2 Metrics

For each subgroup I calculated the difference in median age as the median age in the demonstration data minus the median age in the SF1.

For each subgroup based on geography, size of population, sex and race/ethnicity I calculated a Mean Difference (possibly indicating bias) and Mean Absolute Difference (measure of accuracy). I also tallied the number of times the absolute difference exceeded 1, 2 and 5 years which allows me to created percentage of all cases that exceeds those thresholds.

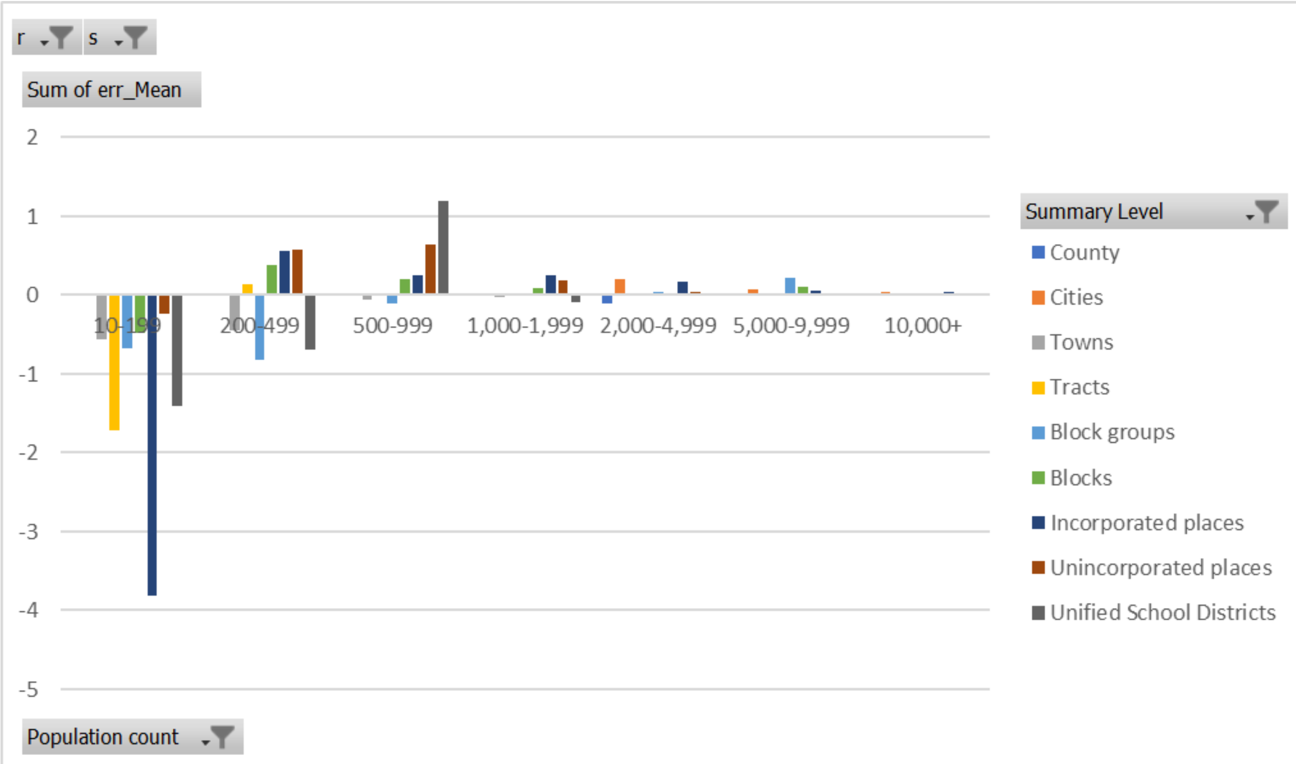
3.4.3 Results

3.4.4 Total population by geography type and size

Number of geographies by type and size

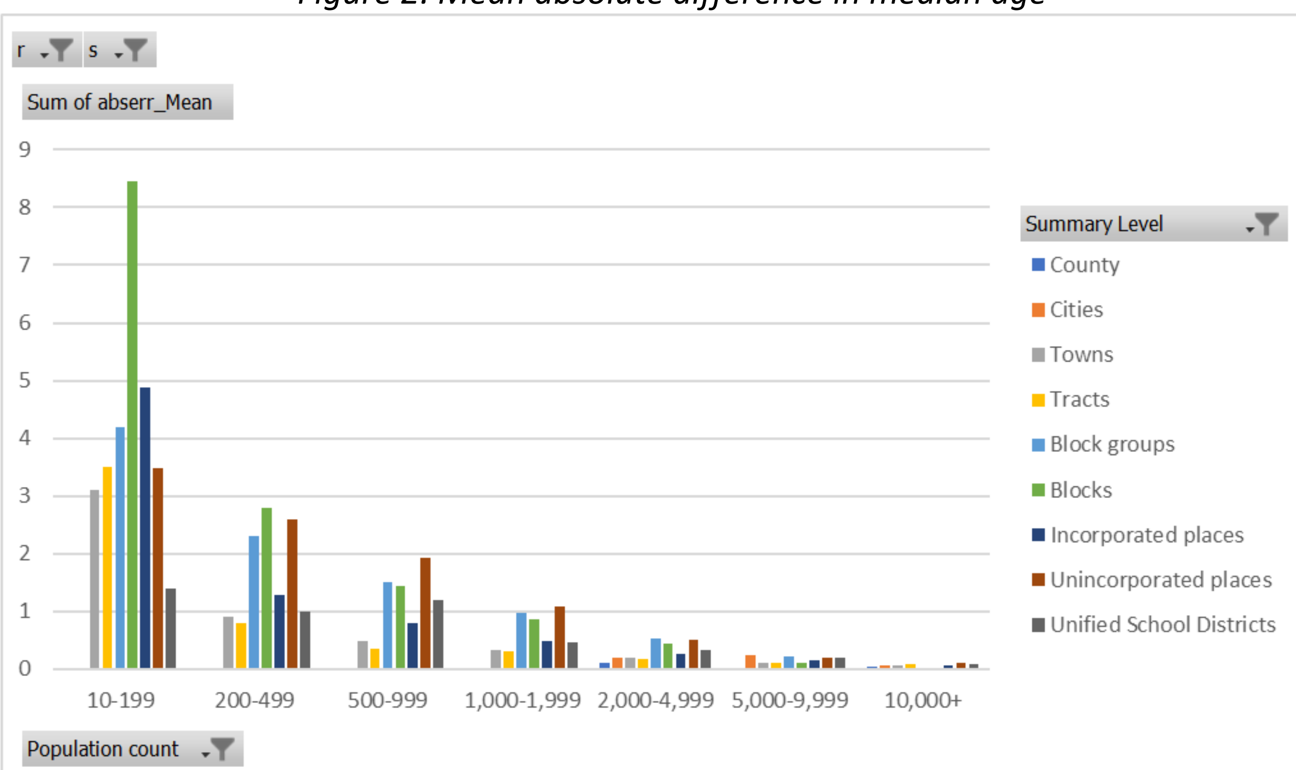
Row Labels	County	Cities	Towns	Tracts	Block groups	Blocks	Incorporated places	Unincorporated places	Unified School Districts
10-199			6	29	60	168169	6	20	1
200-499			25	12	165	14997	67	67	8
500-999			75	15	5097	4475	111	90	5
1,000-1,999			223	535	8336	1097	125	97	18
2,000-4,999	1	2	296	3000	1497	121	139	121	118
5,000-9,999		8	151	1234	14	1	82	72	198
10,000+	61	51	156	30	1		87	103	321

Figure 1: Mean difference in median age



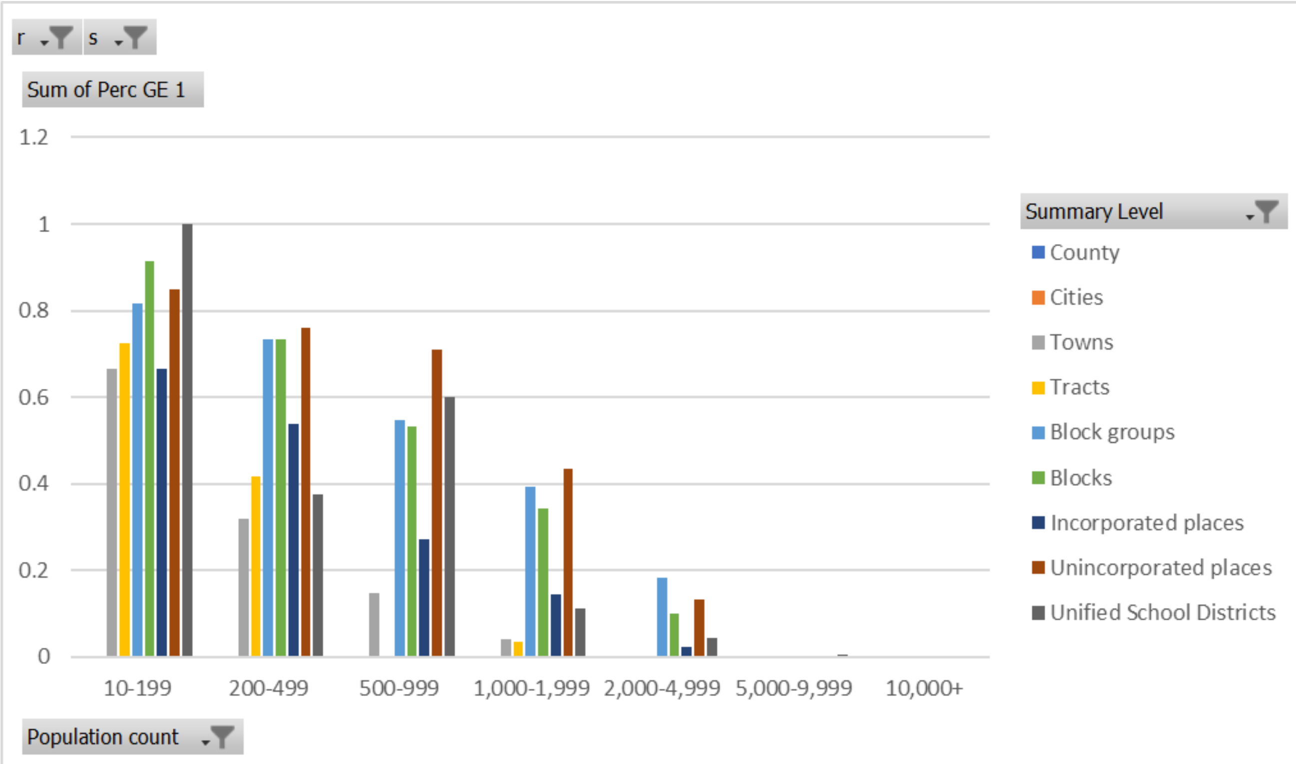
Conclusion: Very small population groups (< 200) seem to have a negative bias

Figure 2: Mean absolute difference in median age



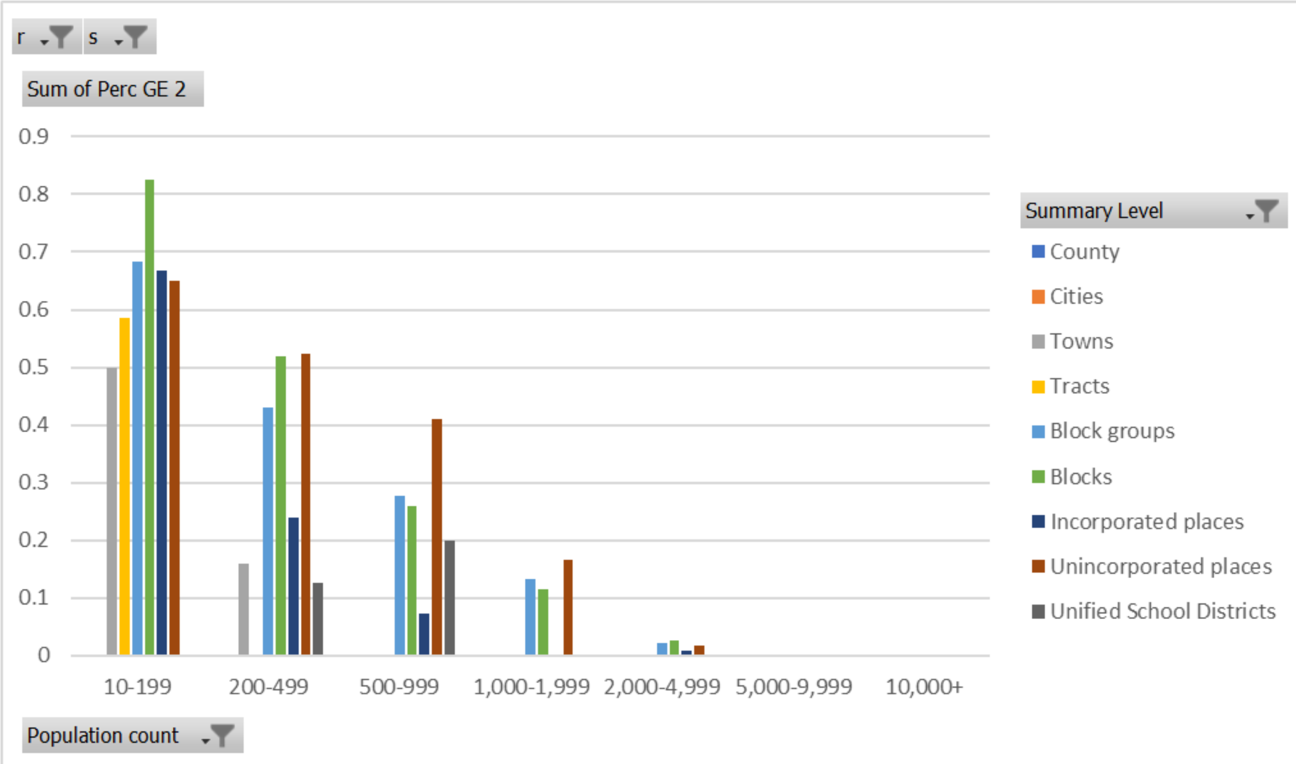
Conclusion: Unincorporated places (CDPs) have larger differences compared to Incorporated places (villages and cities).

Figure 3: Fraction of geographies with an absolute difference in median age of 1 yr or more



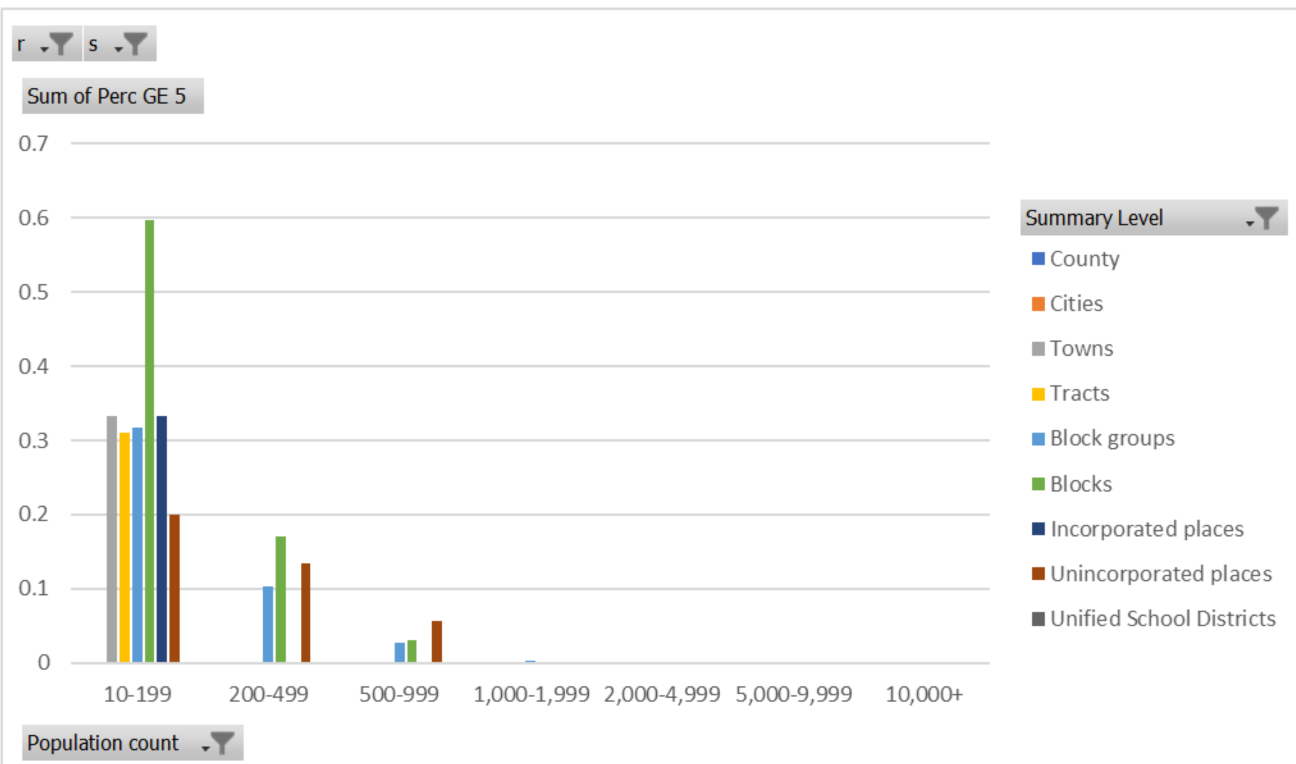
Conclusion: A 1 year difference in median age between the demonstration data and SF1 is quite common for geographies up to 2,000 persons. It is an indication that there is a gap in the cumulative age distributions and this might lead to usability problems.

Figure 4: Fraction of geographies with an absolute difference in median age of 2 yr or more



Conclusion: CDPs, block groups and blocks below 2,000 population often have absolute difference of 2 year or more. This is rare among other geographies of size 500 or more

Figure 5: Fraction of geographies with an absolute difference in median age of 5 yr or more



Conclusion: A 5 year difference between demonstration data and SF1 would certainly lead to usability problems. For geographies above 500 persons, this is rare. Caution is needed for CDP's with populations 500-999.

3.4.5 Mean absolute differences for major race/ethnicity groups

Figure 6: Mean absolute difference in median age for ***Non-Hispanic White Alone*** *populations*

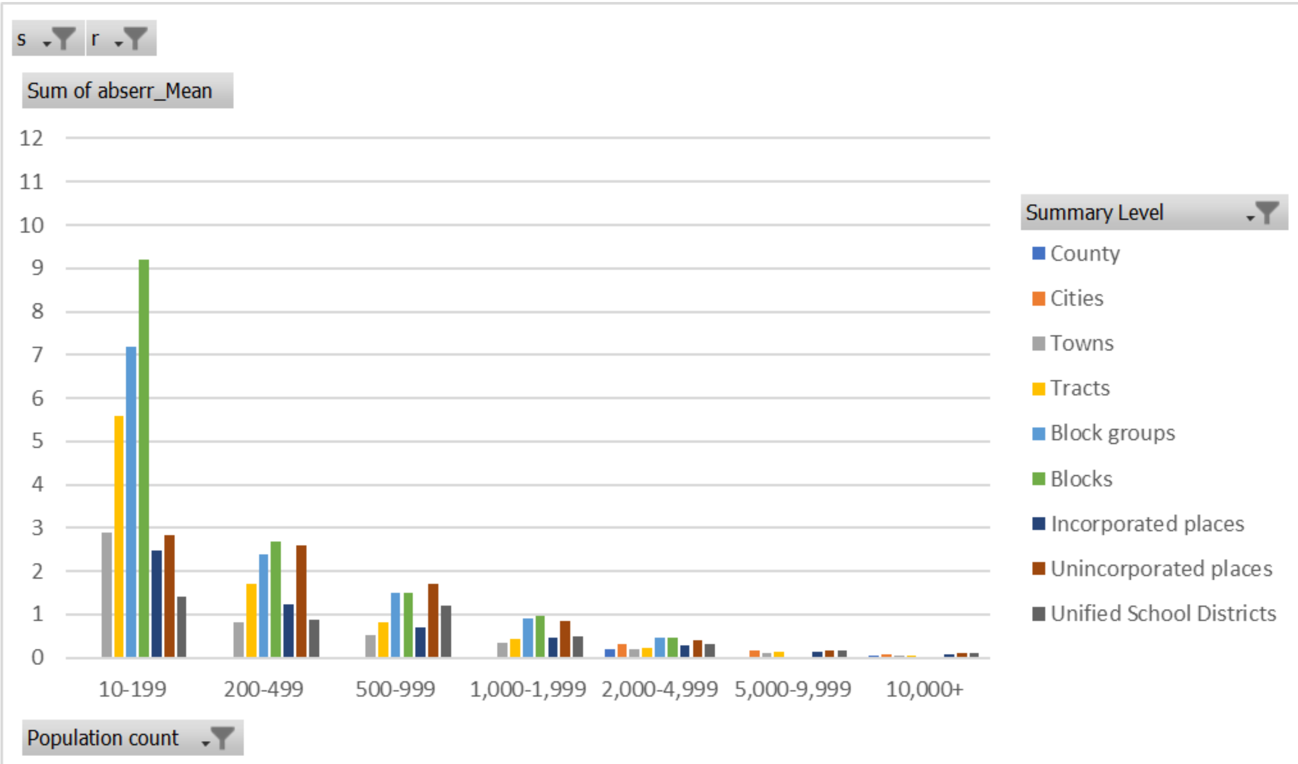


Figure 7: Mean absolute difference in median age for ***Black Alone*** *populations*

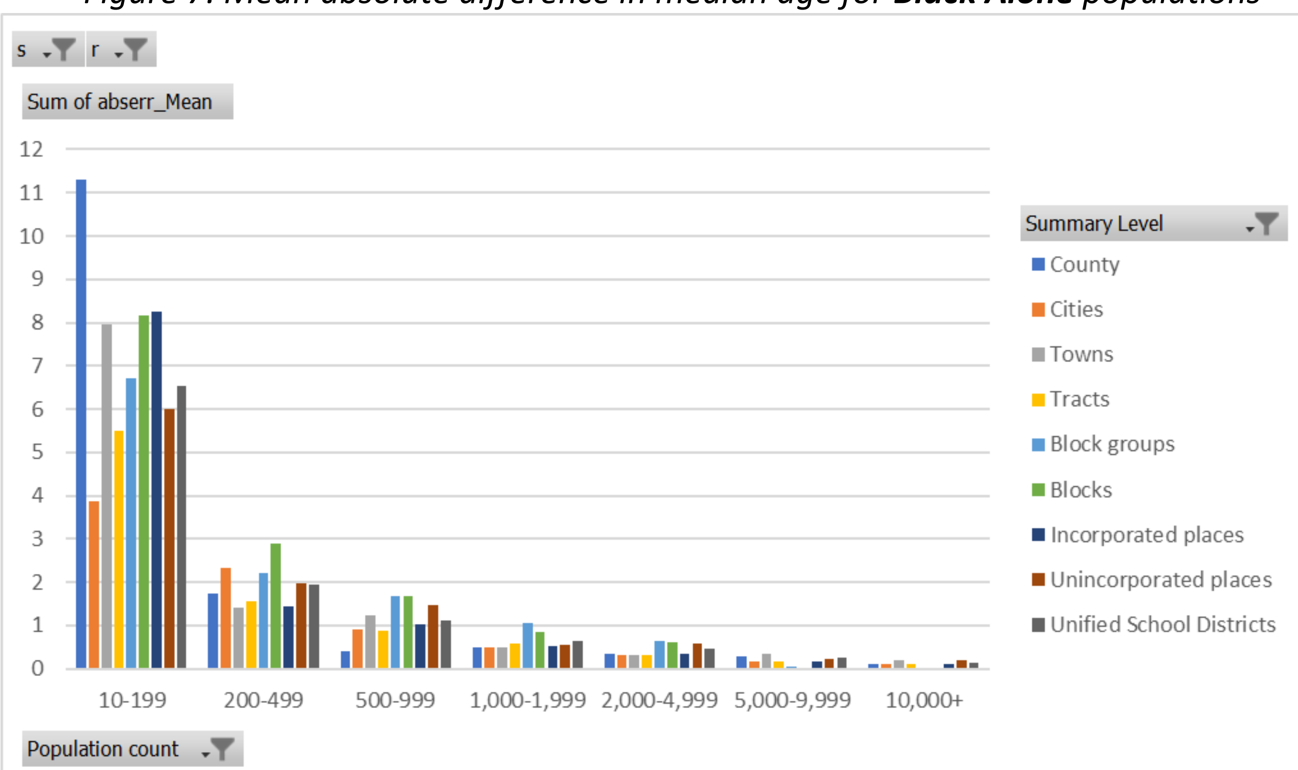
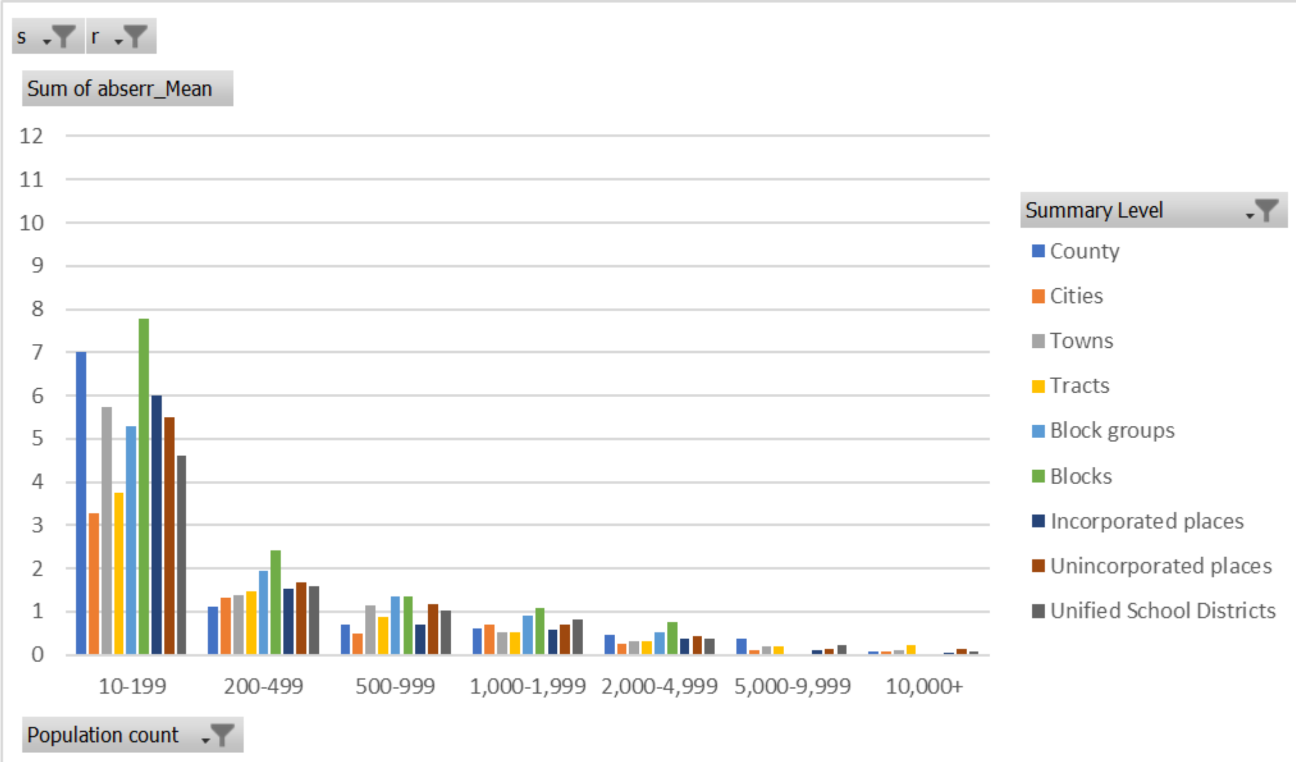


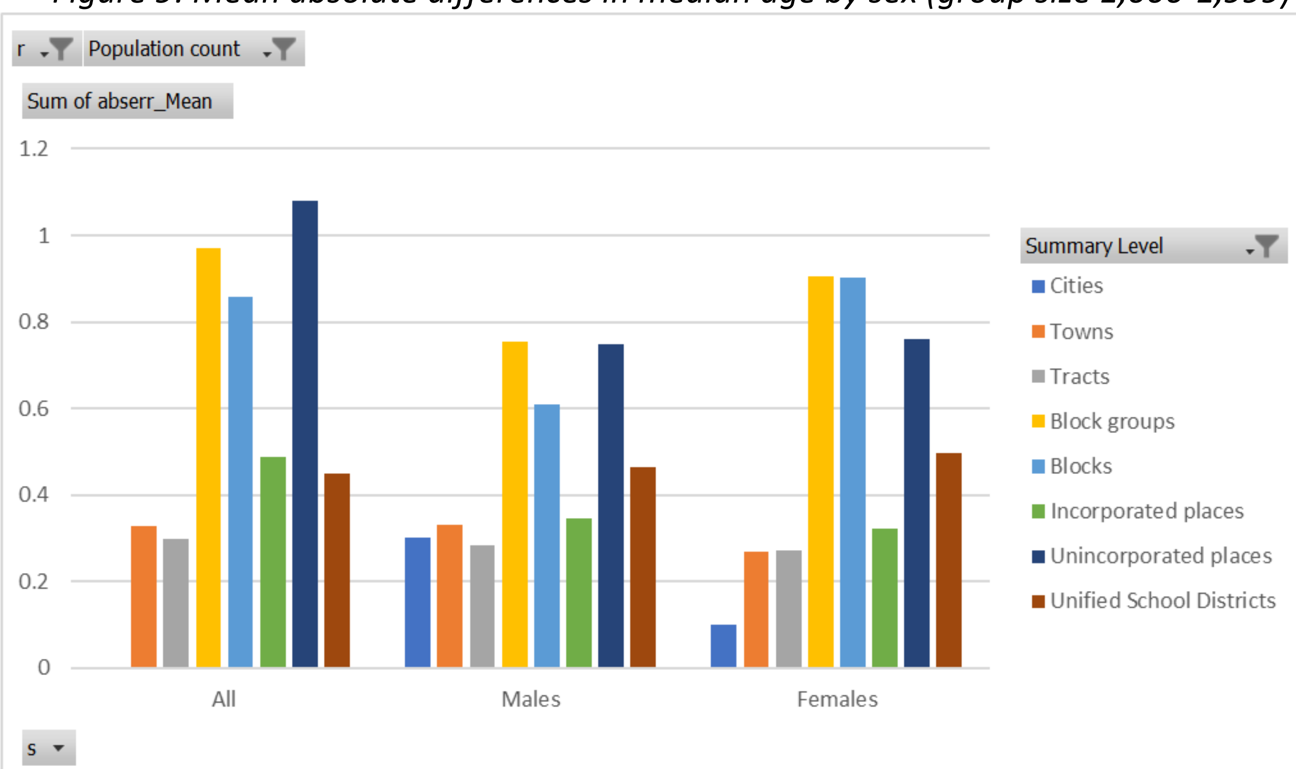
*Figure 8: Mean absolute difference in median age for **Hispanic** populations*



Conclusions: The differences in median age are larger for minority populations Black Alone and Hispanic than they are for Non-Hispanic White Alone populations of similar size. The differences between geographic summary levels seems to be less pronounced for Black Alone and Hispanic population.

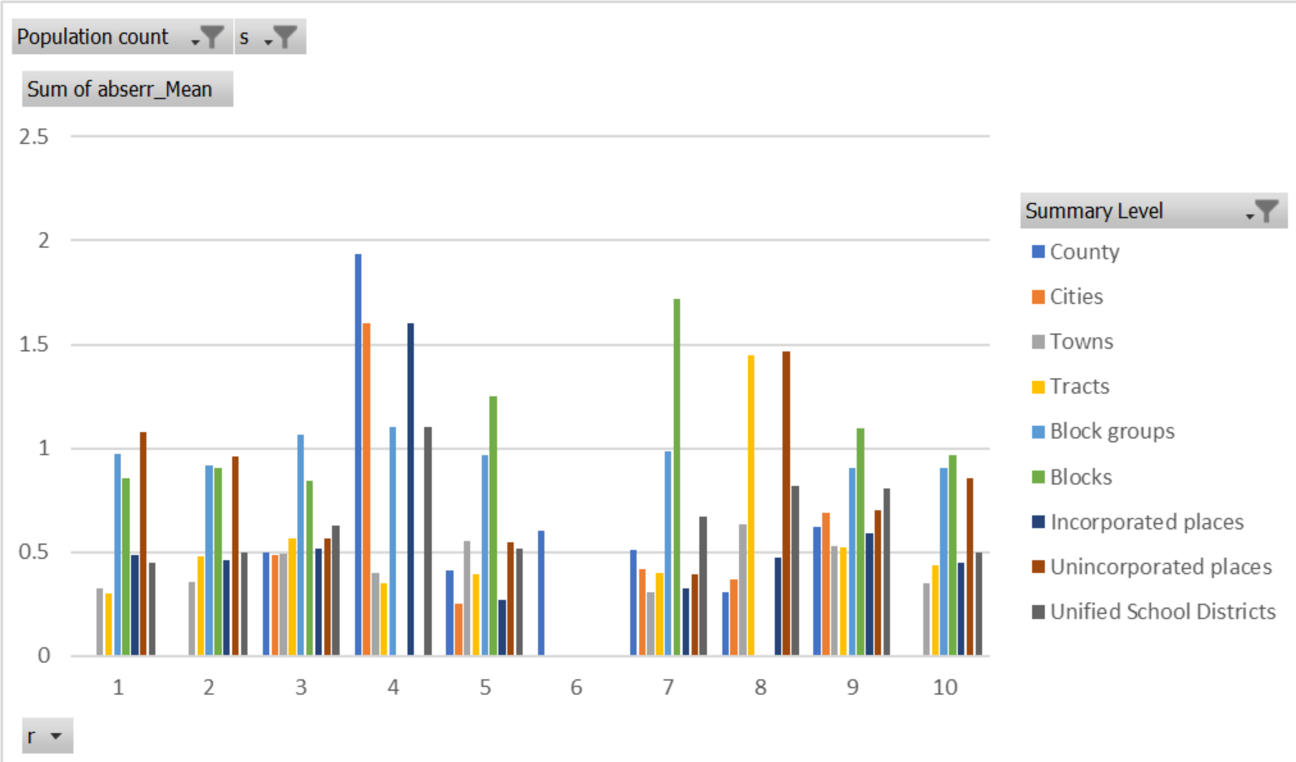
3.4.6 Mean absolute differences for population sizes 1,000 – 1,999

Figure 9: Mean absolute differences in median age by sex (group size 1,000-1,999)



Conclusion: The mean absolute difference in median age for males or females where there are 1,000 to 2,000 males or females is smaller then total populations of 1,000 to 2,000. For some geographic summary levels there is a slight difference between differences in males and females, but I suspect this difference is not significant.

Figure 10: Mean absolute differences in median age by race (group size 1,000-1,999)



Conclusion: The mean absolute difference for groups of two or more races (race code 8) that are 1,000 to 2,000 in size are larger than for other race categories that size. Black (race code 3) and Hispanic (race code 9) hve slightly higher differences than White Alone (race code 2) and NH White Alone (race code 10)

3.5 Age ranges with big differences between the demonstration data and SF1

3.5.1 Method

Data tables used: PCT12, PCT12A-PCT12O (single years of age, iterated by race categories)

Geographies selected: From the national file I selected Counties (SUMLEV = 050) and Places (SUMLEV=160). The PLACECC is used to divide the places in incorporated places (PLACECC starts with a "C") and unincorporated places (PLACECC starts with "U")

The observations are put in 5 bins based on SF1 universe size:

1. 500-999
2. 1,000-1,999
3. 2,000-4,999
4. 5,000-9,999
5. 10,000 plus

For each comparison, I kept tract of the cumulative difference between the demonstration data and SF1:

$$Cumulative\ difference\ at\ age\ x = \sum_{a=0}^{x} Demo(a) - SF1(a)$$

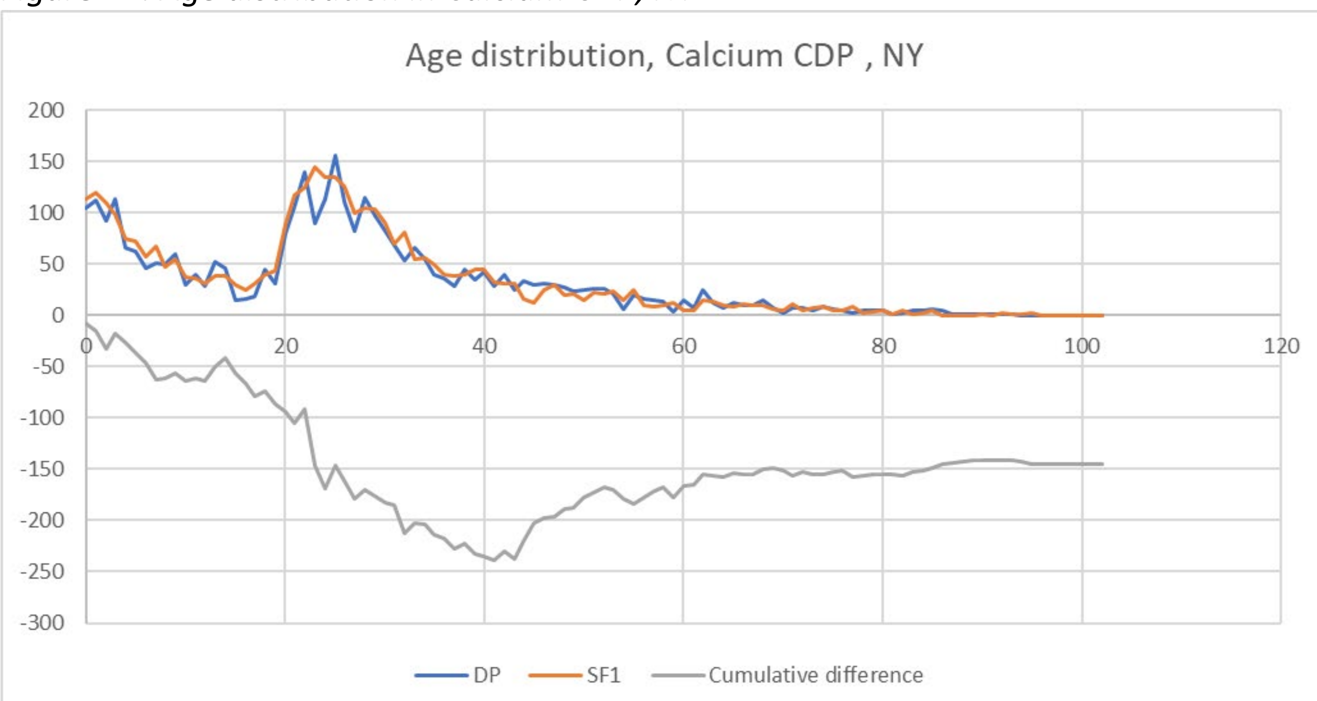
I then looked at the range of these values (max – min) as an indication of the existence of an age range where the cumulative difference went from the minimum value to the maximum value or the other way around. If the range is large than the average error within that age range is significant different from zero.

To determine what a large range is I compared it with two other metrics based on the tables:

- Comparison with the Mean Absolute Error over all ages. If the range is for example 30 times the mean error it indicates that there were at least 30 times differences with the same direction within the a limited age range. After examining the results, I found that this comparison yielded results that were indicative of statistical significance, but not necessarily meant that there would be usability issues
- Comparison with the maximum count in the table. The idea behind this is that when you plot the age distributions and the cumulative differences in a single chart, the plot of the cumulative differences should not have a much bigger range than the range of the age distribution itself.
  NOTE: this method doesn't work well for cases where the population is concentrated in a limited age range

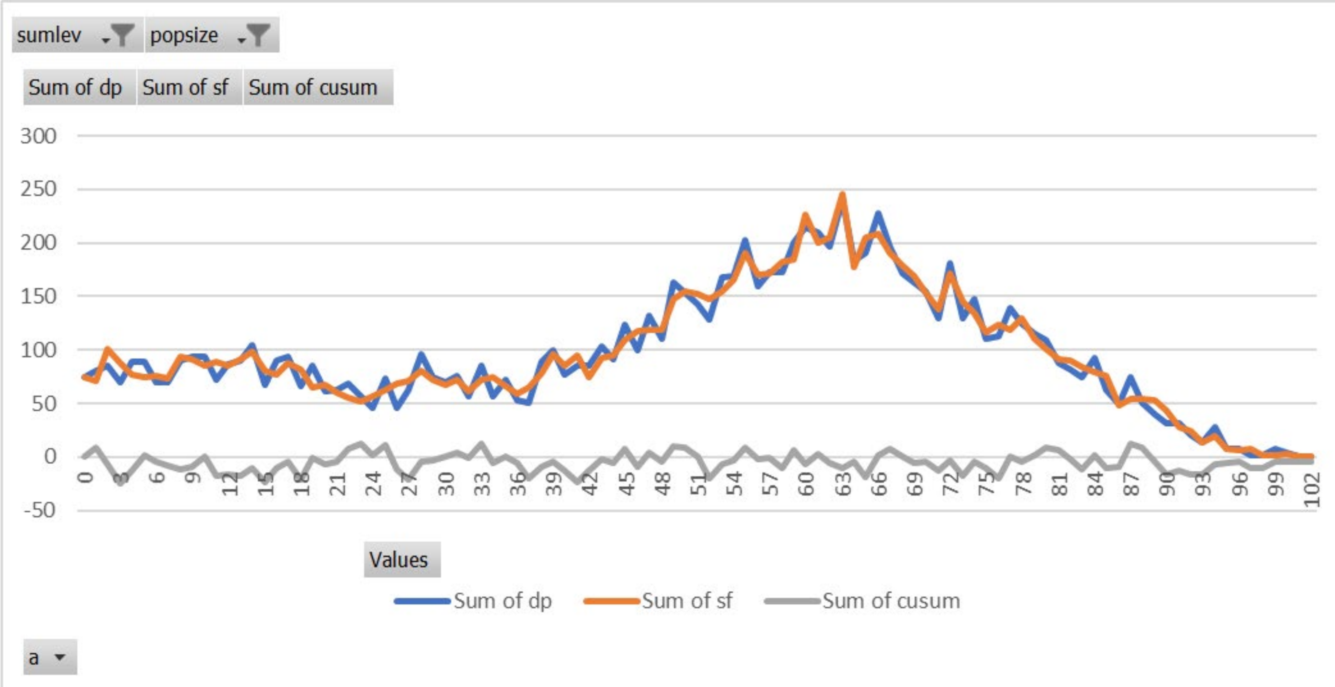
In the example of Calcium CDP I used earlier in this chapter, the cumulative differences almost reaches -250, whereas the peak of the age distribution is close to 150.

Figure 11: Age distribution in Calcium CDP, NY



Underneath is an example of the age distribution of the female population in Llano County, TX

Figure 12: Female age distribution in Llano County, TX



The range of the cumulative differences is 12-(-25)=37 and the maximum count is almost 250. In this case we don't find an age range with a lot of difference between the demonstration data and SF1.

The metric "Range cumulative difference / Max count" doesn't have an easy relation with usability, but it seems that when it exceeds 1.5 it becomes easier to identify an age range where the difference between demonstration data and SF1 could cause usability issues and the larger the value, the easier that gets.

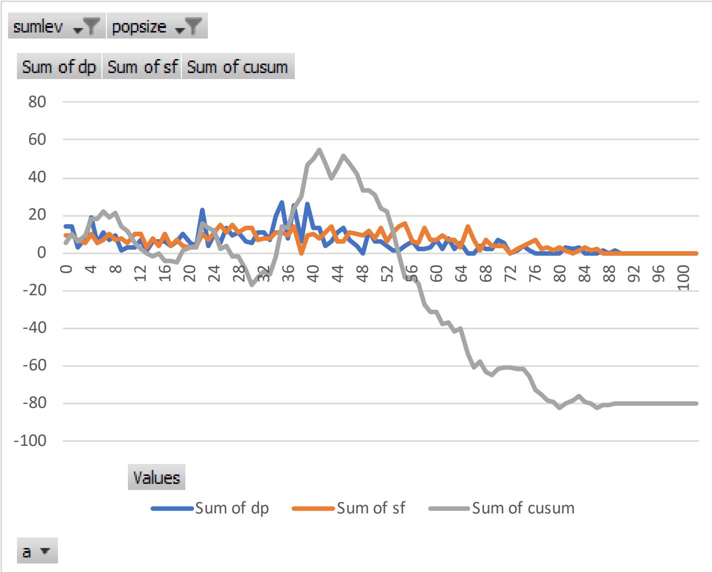
For each of the summary levels and population sizes I measured the percent of observations where the ratio exceeded a variety of thresholds.

			Ratio Range cumulative difference / Max count in SF1				
sumlev	**popsize**	**N**	**>= 1.5**	**>= 2**	**>= 2.5**	**>= 3**	**Maximum**
Counties	500-999	8902	70.2%	42.5%	21.5%	10.0%	7.64
Counties	1,000-1,999	8616	36.2%	12.2%	3.4%	1.1%	5.22
Counties	2000-4,999	11303	4.7%	0.8%	0.1%	0.0%	4.06
Counties	5,000-9,999	8875	0.1%	0.0%	0.0%	0.0%	1.87
Counties	10,000+	23232	0.0%	0.0%	0.0%	0.0%	1.37
Incorporated places	500-999	41860	33.4%	11.8%	3.7%	1.3%	7.00
Incorporated places	1,000-1,999	33571	9.4%	2.0%	0.4%	0.1%	5.25
Incorporated places	2000-4,999	31537	0.8%	0.1%	0.0%	0.0%	3.98
Incorporated places	5,000-9,999	15758	0.03%	0.0%	0.0%	0.0%	1.84
Incorporated places	10,000+	20434	0.01%	0.0%	0.0%	0.0%	1.71
Unincorporated places	500-999	19539	75.8%	51.3%	29.7%	16.7%	8.56
Unincorporated places	1,000-1,999	16307	46.9%	22.5%	9.6%	3.9%	5.81
Unincorporated places	2000-4,999	14322	13.0%	3.4%	0.7%	0.3%	4.39
Unincorporated places	5,000-9,999	5898	0.5%	0.1%	0.0%	0.0%	2.10
Unincorporated places	10,000+	4130	0.02%	0.0%	0.0%	0.0%	1.94

It is clear that geographies with larger populations have less cases where the ratio indicates possible problems. Unincorporated places have many more problems than incorporated places.

A few examples of geographies that caused the maximum ratio:

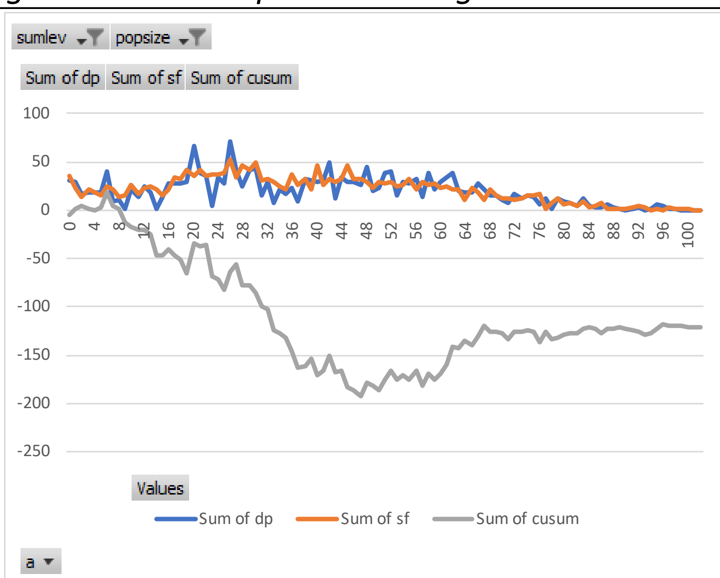
Figure 13: White Alone age distribution in Saxapahaw CDP, NC



SUMLEV	160U				
GEOID	1600000US3759580				
Name	Saxapahaw CDP				
Iteration	A				
Sex	Male				
Total pop	DP	557		Diff	-80
	SF	637			-12.6%
Median age	DP	33.6	SF1	36.7	
Cumulative Error				**Ages**	**45-80**
Minimum		-82		DP	121
Maximum		55		SF1	245
Range		137		Diff	-124
Max SF1		16			-50.6%
	Ratio	8.56			

The male White Alone population in Saxapahaw CDP in North Carolina between the ages and 45 and 80 was 50% lower in the demonstration data compared to SF1

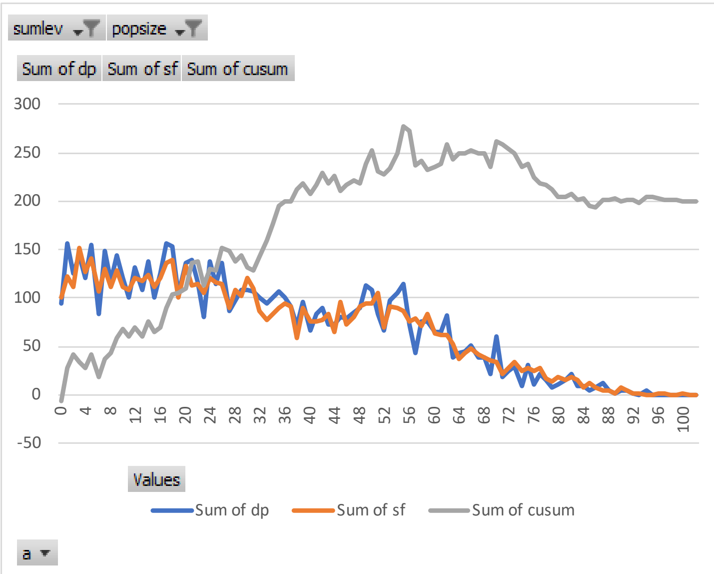
Figure 14: Non Hispanic Black age distribution in New York County, NY



SUMLEV	50				
GEOID	0500000US36061				
Name	New York County				
Iteration	K				
Sex	Total				
Total pop	DP	2023		Diff	-121
	SF	2144			-5.6%
Median age	DP	39.0	SF1	35.6	
Cumulative Error				**Ages**	**5-48**
Minimum		-193		DP	1138
Maximum		18		SF1	1331
Range		211		Diff	-193
Max SF1		52			-14.5%
	Ratio	4.06			

Iteration K is Non-Hispanic Black. The total in New York County, NY [Manhattan] is 5.6% lower in the demonstration data, but between the ages 5 and 48 the difference is 14.5%.

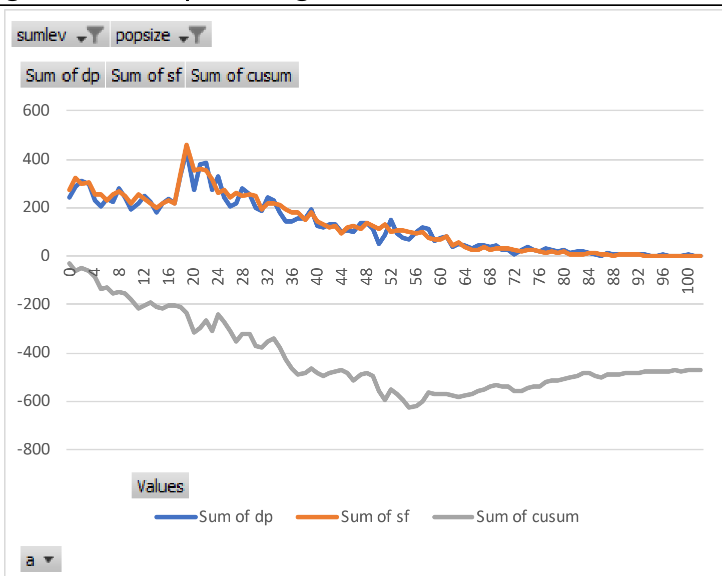
Figure 15: Non-Hispanic Black age distribution in Humboldt County, CA



SUMLEV		50			
GEOID		0500000US06023			
Name		Humboldt County			
Iteration		K			
Sex		Total			
Total pop	DP	7161	Diff		200
	SF	6961			2.9%
Median age	DP	28.0	SF1	28.4	
Cumulative Error			**Ages**	**0-55**	
Minimum		-6	DP		5955
Maximum		277	SF1		5706
Range		283	Diff		249
Max SF1		151			4.4%
	Ratio	1.87			

The metric for the total Non-Hispanic Black population in Humboldt County, CA is 1.87. The median age is very similar between the two data sources. For the ages 0-55 the difference is much bigger than the overall difference (4.4% vs 2.9%), but that might still be usable.

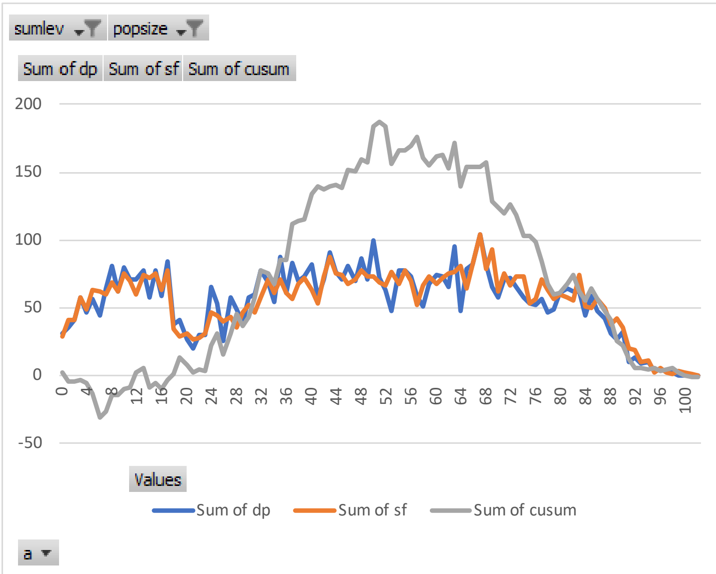
Figure 16: Hispanic age distribution in Humboldt County, CA



SUMLEV		50			
GEOID		0500000US06023			
Name		Humboldt County			
Iteration		H			
Sex		Total			
Total pop	DP	12742	Diff		-469
	SF	13211			-3.6%
Median age	DP	23.0	SF1	22.7	
Cumulative Error			**Ages**	**0-55**	
Minimum		-625	DP		11383
Maximum		-29	SF1		11977
Range		625	Diff		-594
Max SF1		457			-5.0%
	Ratio	1.37			

Humboldt County, CA also had the largest metric value for the over 10,000 category, this time for iteration H (Hispanic population). The difference in the 0-55 yr old population was -5%, which could cause usability problems.

Figure 17: Female age distribution in Viera East CDP, FL



SUMLEV	160U				
GEOID	1600000US1274368				
Name	Viera East CDP				
Iteration	Total				
Sex	Female				
Total pop	DP	5640	Diff		-1
	SF	5641			0.0%
Median age	DP	46.4	SF1	48.4	
Cumulative Error			**Ages**	**5-50**	
Minimum		-31	DP		2810
Maximum		187	SF1		2647
Range		218	Diff		163
Max SF1		104			6.2%
	Ratio	2.10			

The age distribution female population in Viera East CDP, FL has the largest metric value for the 5,000-9,999 population size. Between age 5-50 there was a difference of 6.2% whereas the total population was right on.

4 GEOGRAPHIES WITH IMPOSSIBLE STATISTICS

4.1 RESEARCH QUESTION:

Breaking the connection between households and persons lead to many impossible statistics. How often do they appear for different levels of geography?

4.2 CONCLUSIONS:

- Block level data is full of inconsistencies
- There are a few fields that by definition should be the same in the person file as in the household file (e.g. householders = households, householders living alone = single person households). In this data set, this is very rarely the case. Big differences also occur rather often in sub-county geographies
- The number of older age householders often exceeds the population in that age group
- The number of minority householders often exceeds the population in those race groups. At the block group level the householders often outnumber the population by more than 10 and more than 10%

4.3 METHOD:

The SUMLEV field was used to determine the geographic level. Records with zero population were left out of the analyses. Several inconsistencies were flagged and if the difference exceeded 10 and 10% were flagged as a big error. The percent error is calculated as a difference divided by the average of both observations. Extreme examples were chosen by looking at a combination of the percent error and the size of the population

4.4 MORE HOUSEHOLDS THAN HOUSEHOLD POPULATION

The household population (from table H8) should be larger than the number of occupied houses (from table H3).

Summary level	N	Flagged count	%	Big error count	%
County	62	0		0	
Tract	4870	1	0.02%	0	
Block group	15194	2	0.01%	0	
Blocks	244281	4646	1.9%	84	0.03%
MCD	1010	0		0	
Place	1189	1	0.08%	0	
Unified SD	669	0		0	

Extreme examples:

Block 361019613001000: Total population = 299, Household population = 1, occupied houses = 15

Block 360550094002030: Total population = 140, Household population = 140, occupied houses = 156

4.5 HOUSEHOLD POPULATION WITHOUT OCCUPIED HOUSES

If there is household population (from table H8) than the number of occupied houses (from table H3) should be non-zero.

Summary level	N	Flagged count	%
County	62	0	
Tract	4870	10	0.2%
Block group	15194	12	0.08%
Blocks	244281	16930	6.9%
MCD	1010	2	0.2%
Place	1189	0	
Unified SD	669	0	

Extreme examples:

Block 361031456033001: Total population = 71, Household population = 71, occupied houses = 0 (out of 15 total housing units)

4.6 Householders not equal to households

The population with relationship "householder" (from table P17) should be equal to the number of occupied houses (from table H3).

This is especially important when calculating Persons per Household (PPH) where we often have two different numbers for the denominator.

Summary level	N	Flagged count	%	Big error count	%
County	62	60	96.8%	0	
Tract	4870	4555	93.5%	14	0.3%
Block group	15194	14826	97.6%	1256	8.3%
Blocks	244281	223729	91.6%	47857	19.6%
MCD	1010	964	95.4%	0	
Place	1189	1150	96.7%	86	7.2%
Unified SD	669	655	97.9%	4	0.6%

Extreme examples:

Pleasant Valley CDP: Household population = 1,154,
Householders = 476, Persons per household based on householders = 2.42
Occupied houses = 541, PPH based on occupied houses = 2.13

Blockgroup 3608111551023: Household population = 1,322,
Householders = 529, PPH based on householders = 2.50
Occupied houses = 411, PPH based on occupied houses = 3.22

4.7 HOUSEHOLDERS LIVING ALONE FROM THE PERSON FILE NOT EQUAL TO HOUSEHOLDERS LIVING ALONE FROM THE UNIT FILE

The population with relationship "householder living alone" (from table P17) should be equal to the number of households with household type "Householder living alone" (from table P16).

Summary level	N	Flagged count	%	Big error count	%
County	62	62	100.0%	0	
Tract	4870	4644	95.4%	101	2.1%
Block group	15194	14842	97.7%	7346	48.3%
Blocks	192337	184599	75.6%	34141	14.0%
MCD	1010	956	94.7%	10	1.0%
Place	1189	1139	95.8%	160	13.5%
Unified SD	669	656	98.1%	12	1.8%

Extreme examples:

Crugers CDP: Total population = 1,565, Householders living alone 243 male + 352 female = 595, Household type "householder living alone" = 521

Blockgroup 360610191001: Total population = 1476, Householders living alone 147 male + 234 female = 381, Household type "householder living alone" = 518

Block 360470944013003: Total population = 405, Householders living alone 2 male + 0 female = 2, Household type "householder living alone" = 56

4.8 HOUSEHOLD POPULATION UNDER 18 LESS THEN NUMBER OF HOUSEHOLDS WITH CHILDREN UNDER 18

The household population under age 18 (from table P15) should be at least as large as the number of households with one or more people under 18 (from table P21).

Summary level	N	Flagged count	%	Big error count	%
County	62	0		0	
Tract	4870	14	0.3%	5	0.1%
Block group	15194	196	1.3%	83	0.5%
Blocks	192337	45383	18.6%	3079	1.3%
MCD	1010	3	0.3%	1	0.1%
Place	1189	13	1.1%	3	0.3%
Unified SD	669	2	0.3%	1	0.1%

Extreme examples:

Block group 361190004013: Total population = 670, Household population under 18 = 52, Households with children under 18 = 137

4.9 NOT ENOUGH HOUSEHOLD POPULATION TO FILL THE HOUSEHOLD BY SIZE STATISTICS

One can calculate an under bound for the household population from table H9 (households by size) by multiplying each size category by the size and multiply the 7-or-more category by 7. The household population (table H8) should be larger than this under bound.

One can also calculate what the average household size of the 7 or more category should be to match the household population. Values much larger than 10 are very improbable. These analyses are not part of this feedback.

Summary level	N	Flagged count	%	Big error count	%
County	62	29	46.8%	0	
Tract	4870	1777	36.5%	24	0.5%
Block group	15194	6840	45.0%	847	5.6%
Blocks	192337	102403	41.9%	49766	20.4%
MCD	1010	542	53.7%	16	1.6%
Place	1189	685	57.6%	115	9.7%
Unified SD	669	333	49.8%	7	1.0%

Extreme examples:

Block 361032010014012: Household population = 8, under bound based on housing size = 83 (1*2+1*3+1*4+5*5+7*7) – 15 households in total which is greater than the household population

Delaware County: Household population = 45,556, under bound based on housing size = 45,843 (which is reached when all 190 7+ households have exactly 7 household members)

Thousand Island Park CDP: Household population = 49, under bound based on housing size = 145 (11*1 + 17*2+14*3+4*4+7*5+1*7) – 54 households in total which is greater than the household population

Blockgroup 361190004013: Household population = 670, under bound based on housing size = 1054 (138*1 + 83*2+50*3+38*4+34*5+23*6+20*7)

4.10 MORE HOUSEHOLDERS OF A CERTAIN AGE GROUP THAN POPULATION OF THAT AGE GROUP

The number of people in an age group (from table P12) should greater or be equal to the number of householders in that age group (from table H13)

Geographies without population and without householders in a certain age group are excluded from these analyses.

Figure 18: Share of geographies with number of householders exceeding population by age group

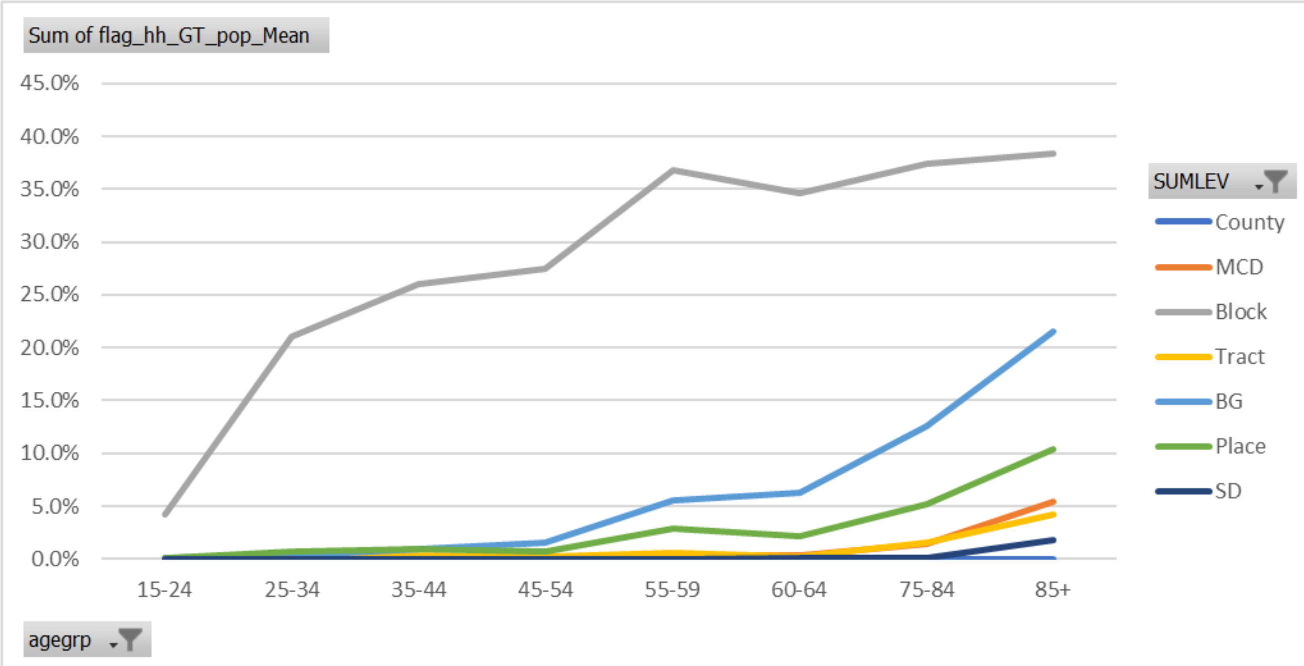
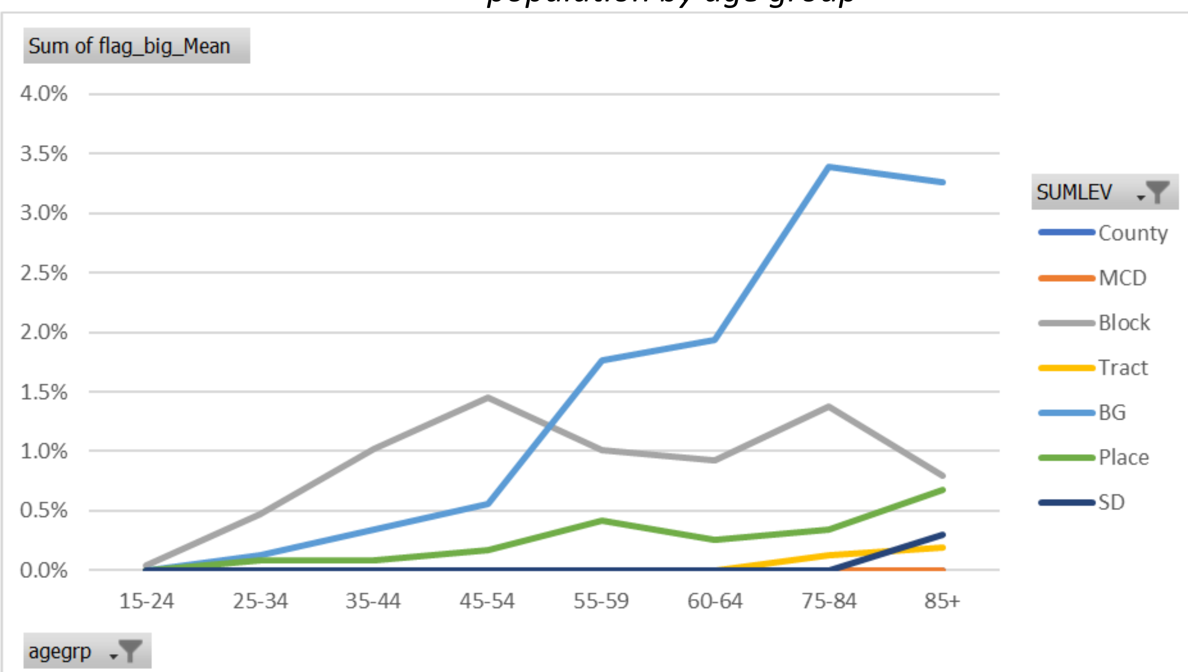


Figure 19: Share of geographies with number of householders greatly exceeding population by age group



Extreme examples:

Blockgroup 360610093005: householders age 60-64 = 53, population age 60-64 = 1

4.11 More householders of a certain race/ethnicity group than population of that same group

The number of people in a race group (from table P5) should greater or be equal to the number of householders in that race group (from table H7)

Geographies without population and without householders in a certain race/ethnicity group are excluded from these analyses.

Figure 20: Share of geographies with number of householders exceeding population by race/ethnicity group

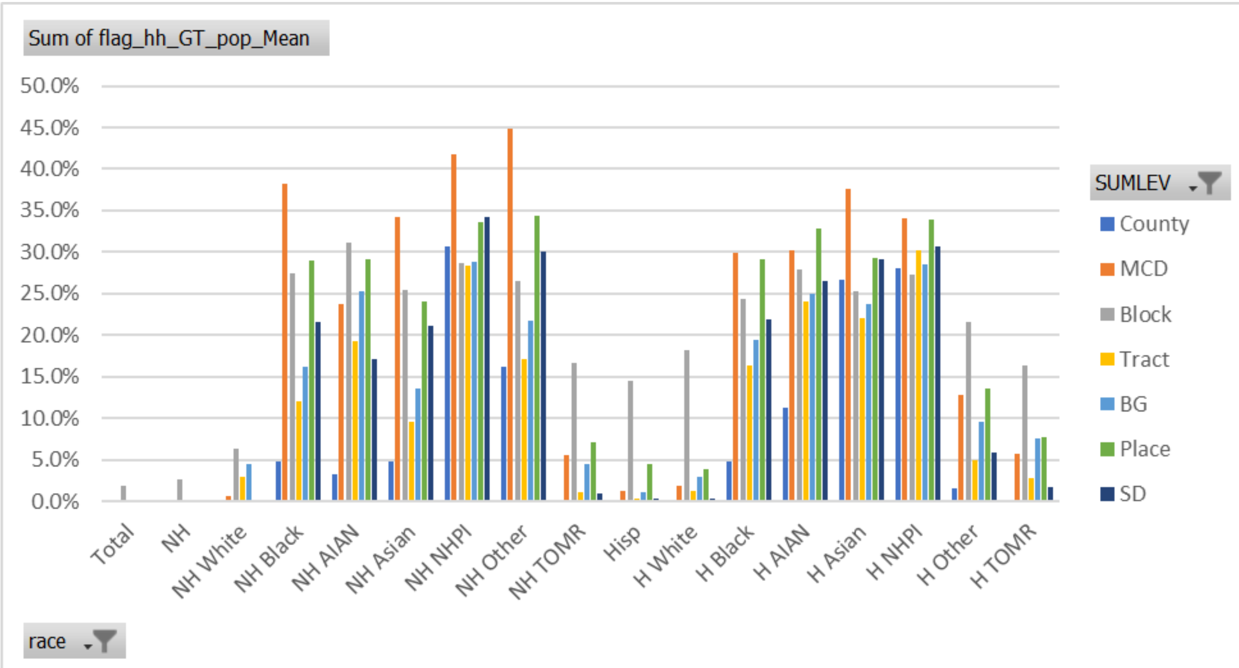
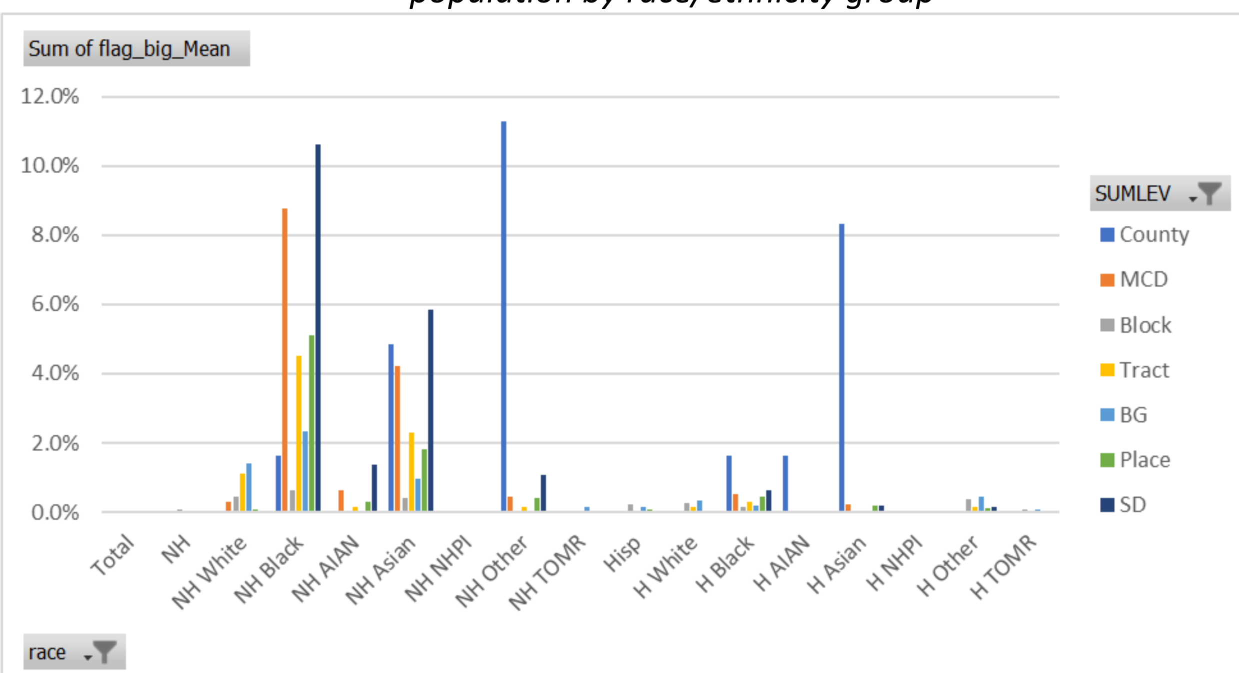


Figure 21: Share of geographies with number of householders greatly exceeding population by race/ethnicity group



Extreme examples:

Blockgroup 360610151002: NH Black Alone householders = 123, NH Black Alone population = 21

5 COMPARING INCORPORATED PLACES, UNINCORPORATED PLACES AND URBAN AREAS

Unincorporated places and urban areas are not on the traditional spine of the Top-Down Algorithm, whereas incorporated places are brought closer to the spine by creating and optimized block group geography.

In this chapter results from incorporated places are compared with unincorporated places and urban areas.

5.1 RESEARCH QUESTION:

Is there much difference in error metrics between incorporated places, unincorporated places and urban areas.

5.2 CONCLUSIONS:

- Severe errors are much more frequent for unincorporated places than for incorporated places for similar size. Urban areas look slightly better than unincorporated places, probably benefitting from the accuracy of the incorporated place at the heart of the urban area.
- Big errors are too common for household type: “Male Householder, no spouse present” and “householder not living alone”. Even in incorporated places with more than 5,000 residents, differences of more than 10% were quite common.
- Larger places (regardless of type of place) showed many severe errors for the number of larger households.

5.3 METHOD:

From the national Summary File (Demonstration data and SF1 – Urban/Rural Update) I extracted geographies with SUMLEV = 160 (places) and SUMLEV = 400. For the places I used the first character of PLACECC variable to split the places in Incorporated places (PLACECC starts with “C”) and Unincorporated places (starts with “U”). All places with other place types were ignored, like military places.

I further restricted my observations to those where at least 80% of the population resided in households.

The observations are put in 5 bins based on SF1 universe size:

1. 500-999
2. 1,000-1,999
3. 2,000-4,999 (Urban areas are by definition larger then 2,500)
4. 5,000-9,999
5. 10,000 plus

I chose three tables for further analyses.

- P12: Population by age and sex. I created 5 year age groups and only looked at the males+females
- P16: Household type
- H13: Household size

For each field I calculated an absolute percent error as the absolute difference between the demonstration data divided by the average of demonstration data and SF1: |DP – SF1| / 0.5 * (DP + SF1). If this absolute percent error was more or equal to 10% and the absolute difference was also more or equal to 10 then I flagged it as a big error.

The tables in this section reflect the share of the observations that showed big errors. The color is a visual aid, the darker the higher the share. All tables use the same coloring scheme with the darkest red reflecting 50% or more observations with big errors.

5.4 Results

Table 1: Share of observations with big errors in table P12: Population by age and sex

	Incorporated places						Unincorporated places						Urban areas		
	0-499	500-999	1,000-1,999	2,000-4,999	5,000-9,999	10,000+	0-499	500-999	1,000-1,999	2,000-4,999	5,000-9,999	10,000+	2,500-4,999	5,000-9,999	10,000+
Total	0.3%	0.0%	0.0%	0.0%	0.0%	0.0%	1.2%	0.3%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
0-4	0.3%	1.3%	2.3%	0.4%	0.0%	0.0%	8.1%	26.1%	29.9%	18.8%	3.0%	0.5%	7.7%	2.1%	0.1%
5-9	0.1%	0.9%	1.3%	0.1%	0.0%	0.0%	8.4%	27.2%	30.8%	15.7%	2.3%	0.4%	7.6%	1.4%	0.1%
10-14	0.2%	1.1%	1.8%	0.2%	0.0%	0.0%	10.0%	27.3%	29.4%	15.0%	2.2%	0.3%	8.6%	1.9%	0.1%
15-19	0.8%	3.2%	5.7%	1.1%	0.0%	0.0%	11.0%	32.1%	36.2%	18.2%	3.2%	0.8%	10.5%	1.9%	0.0%
20-24	0.1%	0.5%	1.6%	0.5%	0.0%	0.0%	8.2%	27.9%	33.3%	23.9%	5.4%	0.7%	11.1%	3.0%	0.1%
25-29	0.1%	0.8%	1.3%	0.4%	0.1%	0.0%	8.9%	28.4%	33.2%	23.3%	6.4%	0.7%	9.0%	2.6%	0.1%
30-34	0.1%	0.6%	1.4%	0.3%	0.0%	0.0%	9.6%	27.3%	34.1%	23.1%	5.5%	0.7%	10.4%	2.8%	0.2%
35-39	0.3%	1.4%	2.0%	0.2%	0.0%	0.0%	11.4%	33.1%	38.1%	22.5%	5.4%	0.8%	11.6%	3.2%	0.1%
40-44	0.3%	1.7%	2.3%	0.5%	0.0%	0.0%	13.8%	32.6%	40.0%	21.9%	3.5%	0.8%	12.2%	2.5%	0.2%
45-49	0.3%	1.2%	1.6%	0.1%	0.0%	0.0%	15.1%	38.4%	41.0%	18.2%	3.9%	0.1%	10.1%	1.7%	0.3%
50-54	0.3%	1.3%	1.6%	0.1%	0.0%	0.0%	17.1%	40.4%	39.2%	18.9%	3.4%	0.1%	10.0%	2.1%	0.3%
55-59	0.2%	0.9%	1.6%	0.2%	0.0%	0.0%	16.2%	39.7%	40.3%	21.4%	4.1%	0.1%	11.3%	4.6%	0.2%
60-64	0.5%	2.4%	4.0%	0.5%	0.0%	0.0%	15.7%	36.4%	39.0%	24.3%	5.9%	0.7%	13.2%	5.4%	0.4%
65-69	0.3%	1.7%	3.4%	1.7%	0.1%	0.0%	11.0%	29.2%	35.9%	27.5%	6.9%	0.8%	17.1%	6.8%	0.9%
70-74	0.0%	0.3%	0.6%	0.8%	0.1%	0.0%	7.4%	21.1%	31.4%	28.4%	10.0%	1.9%	16.4%	7.8%	0.9%
75-79	0.0%	0.3%	0.4%	1.0%	0.4%	0.0%	4.0%	16.2%	22.8%	24.8%	15.8%	3.5%	17.3%	9.2%	1.2%
80-84	0.0%	0.1%	0.6%	0.8%	0.6%	0.1%	1.7%	8.5%	17.7%	19.7%	14.6%	6.4%	14.4%	10.4%	1.5%
85+	0.2%	1.4%	3.2%	4.4%	4.3%	1.0%	1.1%	7.7%	15.5%	22.1%	19.1%	11.3%	15.5%	11.4%	2.3%

Table 2: Share of observations with big errors in table P16: Household type

	Incorporated places						Unincorporated places						Urban areas		
Row Labels	0-499	500-999	1,000-1,999	2,000-4,999	5,000-9,999	10,000+	0-499	500-999	1,000-1,999	2,000-4,999	5,000-9,999	10,000+	2,500-4,999	5,000-9,999	10,000+
Total:	1.4%	0.2%	0.0%	0.0%	0.0%	0.0%	4.0%	0.7%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Family households:	5.8%	2.0%	0.2%	0.1%	0.0%	0.0%	22.9%	18.9%	6.6%	0.9%	0.1%	0.0%	0.4%	0.0%	0.0%
Married couple family	0.4%	0.5%	0.0%	0.0%	0.0%	0.0%	26.1%	37.1%	23.0%	8.5%	1.6%	0.2%	2.6%	0.1%	0.1%
Other family:	4.6%	12.3%	14.8%	7.4%	1.5%	0.2%	11.2%	40.9%	51.9%	42.5%	27.0%	10.1%	19.7%	10.5%	1.8%
Male householder, no spous	0.7%	3.5%	7.6%	16.9%	19.7%	4.9%	1.3%	11.7%	26.6%	40.8%	48.9%	32.6%	30.0%	30.8%	9.1%
Female householder, no spo	0.8%	5.3%	11.0%	7.3%	1.8%	0.2%	6.4%	30.5%	41.9%	42.1%	26.2%	9.5%	22.0%	13.1%	2.1%
Nonfamily households:	3.3%	6.0%	2.7%	1.0%	0.1%	0.0%	22.0%	43.6%	35.0%	17.1%	5.2%	1.3%	5.4%	1.6%	0.2%
Householder living alone	2.2%	4.4%	2.1%	0.4%	0.3%	0.0%	19.7%	41.7%	36.4%	18.9%	6.7%	1.2%	6.0%	2.4%	0.3%
Householder not living alone	0.2%	2.7%	8.4%	20.5%	24.4%	7.3%	2.1%	13.1%	22.0%	35.2%	30.3%	18.1%	30.5%	32.6%	9.4%

Table 3: Share of observations with big errors in table H13: Household size

	Incorporated places						Unincorporated places						Urban areas		
	0-499	500-999	1,000-1,999	2,000-4,999	5,000-9,999	10,000+	0-499	500-999	1,000-1,999	2,000-4,999	5,000-9,999	10,000+	2,500-4,999	5,000-9,999	10,000+
Total:	1.4%	0.2%	0.0%	0.0%	0.0%	0.0%	4.0%	0.7%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
1-person household	2.2%	4.4%	2.1%	0.4%	0.3%	0.0%	19.7%	41.7%	36.4%	18.9%	6.7%	1.2%	6.0%	2.4%	0.3%
2-person household	5.9%	15.7%	6.1%	1.2%	0.0%	0.0%	22.7%	42.2%	27.3%	11.0%	1.7%	0.7%	5.8%	0.7%	0.4%
3-person household	3.0%	15.0%	24.4%	13.2%	3.5%	0.2%	11.2%	34.8%	41.9%	23.2%	4.3%	0.5%	21.9%	7.1%	0.4%
4-person household	2.4%	11.9%	21.6%	18.1%	4.7%	0.4%	8.6%	30.2%	40.9%	28.7%	7.9%	1.5%	27.6%	12.4%	1.1%
5-person household	1.3%	9.4%	18.1%	29.6%	24.5%	4.4%	3.4%	17.0%	27.6%	33.7%	22.4%	6.1%	36.2%	28.6%	6.6%
6-person household	0.3%	3.9%	11.8%	25.7%	41.0%	29.6%	0.5%	3.4%	10.1%	21.6%	31.5%	27.9%	27.9%	41.2%	27.0%
7-or-more-person ho	0.2%	2.9%	8.3%	21.1%	39.8%	44.2%	0.4%	3.1%	7.7%	17.3%	30.1%	39.0%	20.8%	37.0%	38.3%

6 RACE/ETHNICITY OF HOUSEHOLDER

6.1 RESEARCH QUESTIONS:

- Is there significant difference between the race distribution of the householders in SF1 and in the DHC?
- What are the differences in headship and homeownership rates by race?
- What are the differences in race/ethnicity of householders of three generational households?

The Census Bureau presentations for FSCPE and NAC indicated that these questions were investigated at the Census Bureau as well. I think improved accuracy within these tables is very important for housing affordability and access.

Assuming that improvements that were mentioned in the NAC webinar are being implemented, I only show limited results from our own analysis on this topic for the first two questions.

6.2 CONCLUSIONS

The diversity of householders in the demonstration data is often higher than for the same geography in SF1. This is especially true if there was little diversity in SF1.

The calculation of headship rates require data from the persons and data from the households which is problematic and often leads to impossible results.

Diversity among householders in households with three or more generations is rather different between the demonstration data and SF1

6.3 RACE/ETHNICITY DISTRIBUTION OF HOUSEHOLDERS (TABLE H7)

I selected tracts in New York with householders in the demonstration data and in SF1. I than calculated a diversity index based on the 7 Non-Hispanic race groups and the 7 Hispanic race-groups for the demonstration data and for SF1 and looked at the difference between those two.

The tracts were split in three groups based on the SF1 diversity index. The diversity index can be seen as the probability index that two random people are of the same race. A low diversity index means a very homogeneous group and a high index is observed in very diverse geographies.

The mean and mean absolute differences are calculated as well as an indicator which of the two sources indicated more diversity.

Table 4: Difference in tract level diversity between demonstration data and SF1, grouped by SF1 diversity

		Difference between DP and SF1			
Diversity Index value	N	Mean	Mean Absolute	P95 absolute	Prob(DP > SF1)
Low diversity (DI <= 1/3)	2426	2.5%	3.3%	10.0%	70.1%
Medium diversity (1/3 < DI <= 2/3)	1527	0.9%	2.1%	6.3%	62.3%
High diversity (DI > 2/3)	894	0.2%	0.8%	2.6%	53.8%

The mean difference was greater than zero in all three groups indicating that the demonstration data had more diversity among householders than SF1. This is confirmed with the share of observations where the demonstration data had more diversity than SF1.

Another way of looking at this table is by setting thresholds, just like the Census Bureau did for PL95. For 95% of observations the absolute percent error for the largest race of householder category should be less than 5%.

I analyzed the tracts and looked for the minimal population size where the tracts with a larger population had an absolute error of less then 5% in the largest race group. For the tracts in New York State this threshold for the tracts was 2,717 people.

- The absolute percent error for the largest race group within the 3,543 tracts with a population > 2717 had an average of 1.8% and exceeded 5% in 177 tracts (5% of N). The mean percent error was -0.75% indicating a slight negative bias.
- The absolute percent error for the largest race group within the 1,312 tracts with a population <= 2717 had an average of 7.3% and exceeded 5% in 395 tracts (30% of N). The mean percent error was -6.0% indicating a significant negative bias.

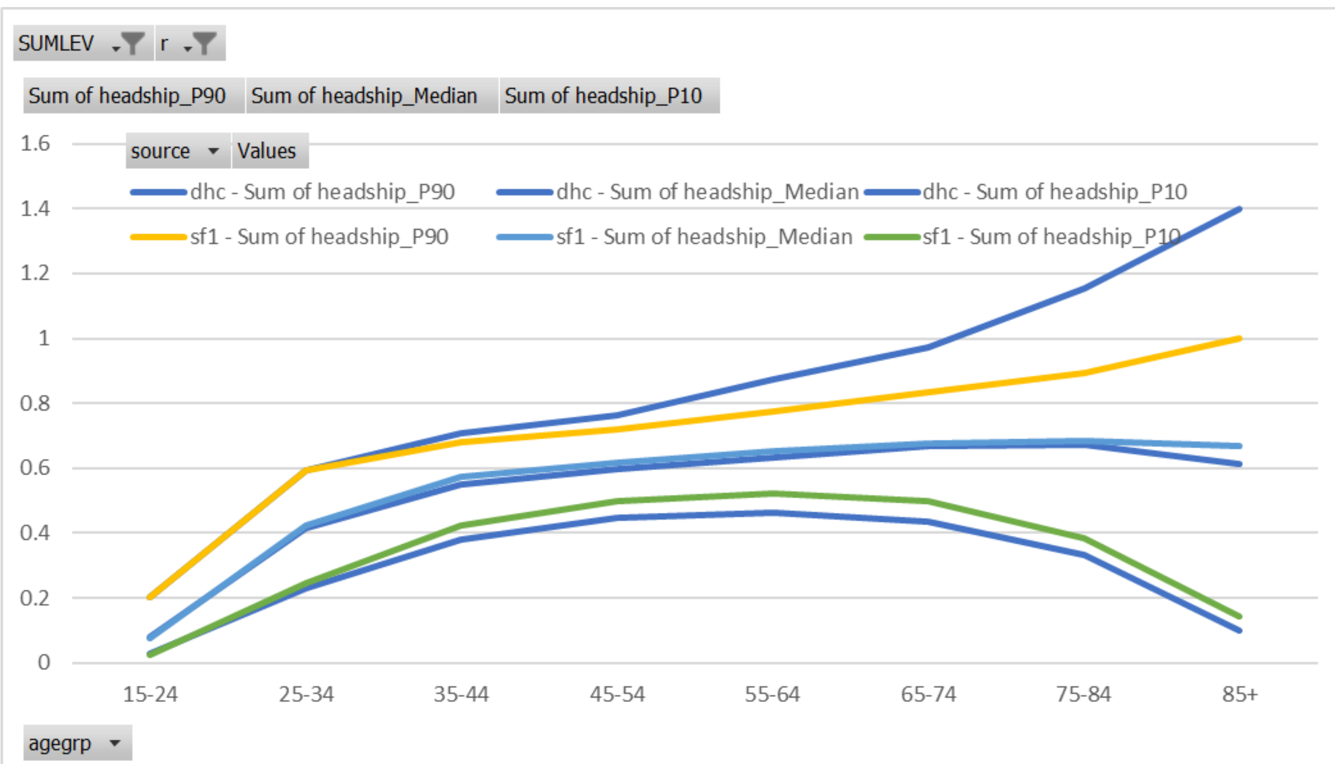
6.4 HEADSHIP AND HOMEOWNERSHIP RATES

I compared the distributions of headship rates (householders in a group divided by household population in that group) and of homeownership rates (homeowners in a group divided by householders in that group) for the several race groups and age groups as presented in table H13 and its iterations.

I compared the 10th percentile, the median and the 90th percentile for each of rates for different levels of geography (with at least 200 householders in the group) between the demonstration data and SF1.

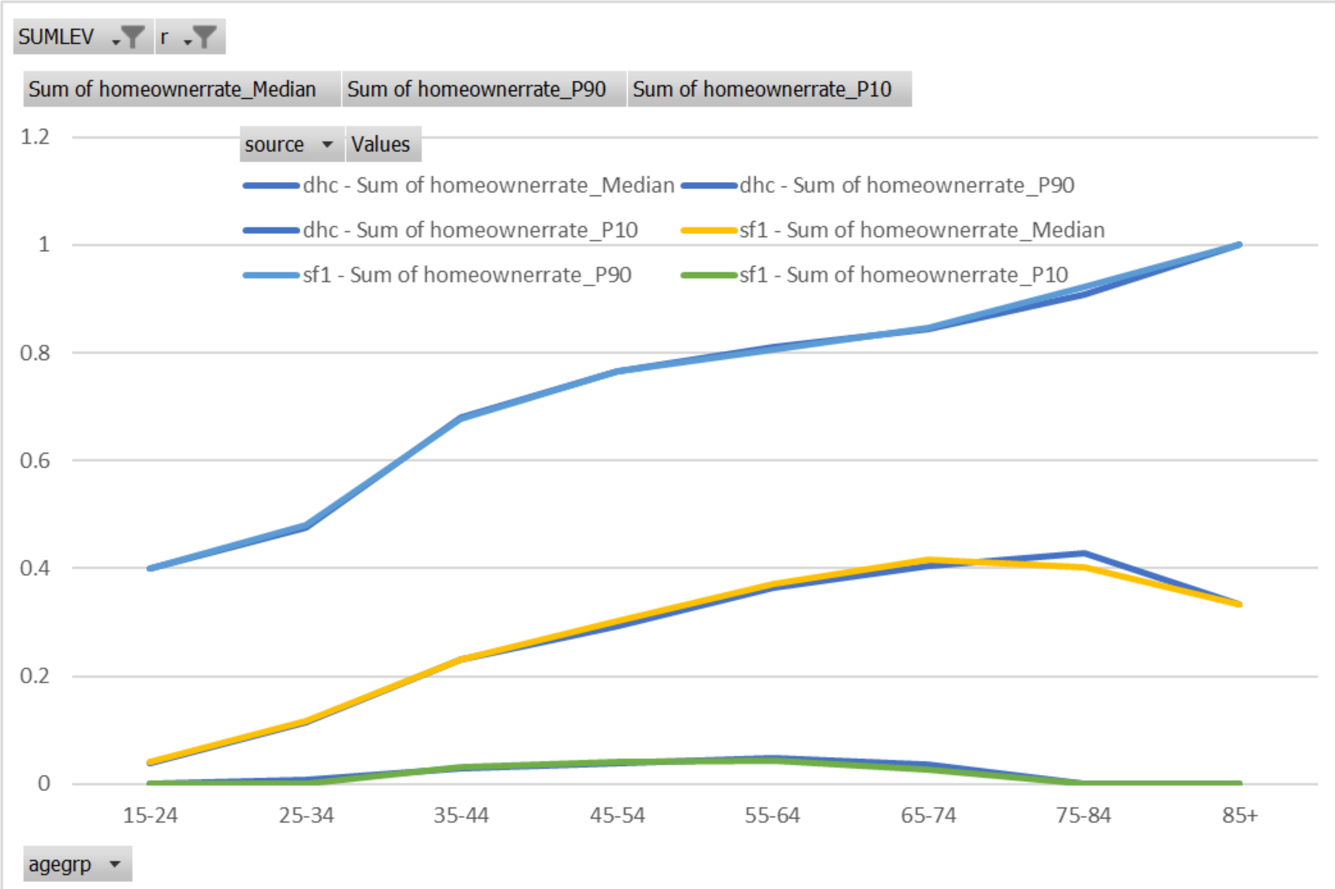
I present here only tract level results for Black householders as they are representative of these analysis.

Figure 22: Distribution of tract level Black Alone headship rates from demonstration data and SF1



The median headship rates are not very different, except for the 85+ population, but the P10 and especially P90 of the headship rate distribution show differences. Theoretically headship rates can not exceed 1, but because population and households are treated independently can be greater than 1 in the demonstration data. For higher ages this happened in more than 10% of tracts and this caused the P90 to exceed 1 in the demonstration data.

Figure 23: Distribution of tract level Black Alone homeownership rates from demonstration data and SF1



The distribution of homeownership rates at the tract level looks very similar between demonstration data and SF1. Please keep in mind that this is not an analysis of the differences at the tract level, only a look at distribution as a result of the aggregate.

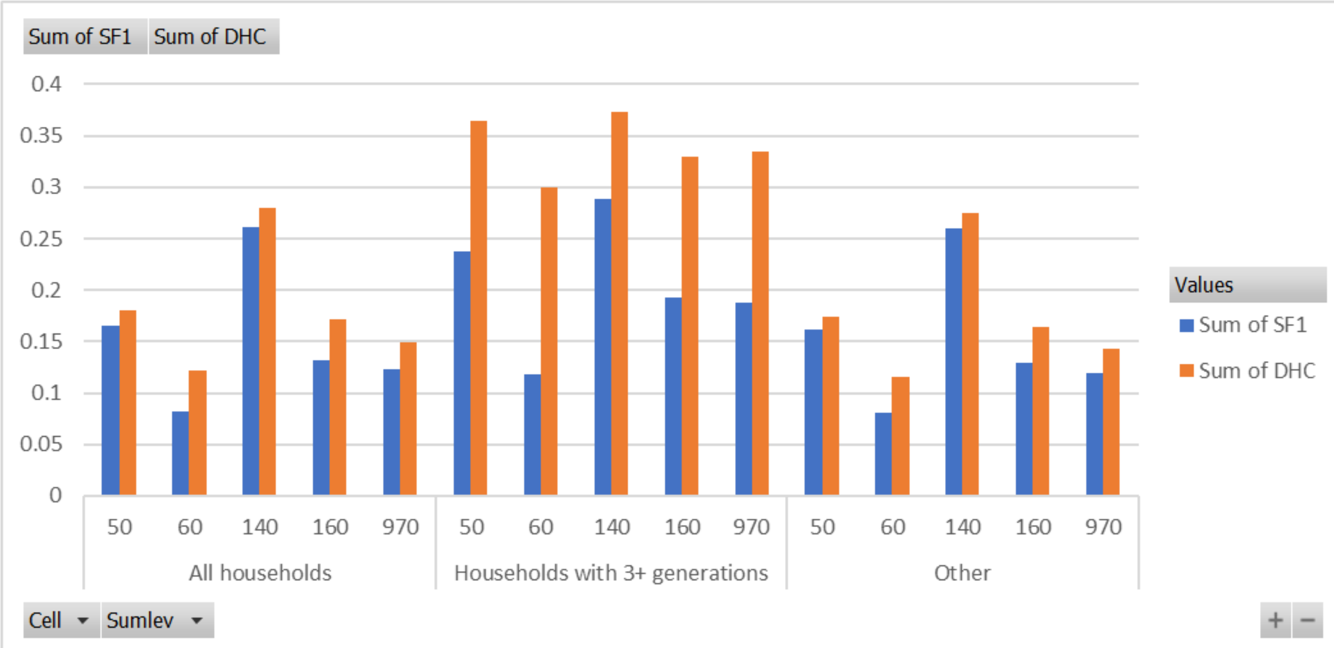
6.5 Householders of three generational households

For this analysis we looked at table PCT14: PRESENCE OF MULTIGENERATIONAL HOUSEHOLDS and it's iterations by the major race groups.

Geographies under consideration are Counties (050), SubCounties (060), tracts (140), places (160) and unified school districts (970) in New York State.

Based on table iterations A through H, I calculated separate diversity indices for all householders and for householders in households with or without the presence of three generations. This diversity index lost a bit of practical interpretation because the groups under consideration are not mutual exclusive, but differences in these diversity indices are still indicative of problems.

Figure 24: Diversity index for householders of households with or without three or more generations



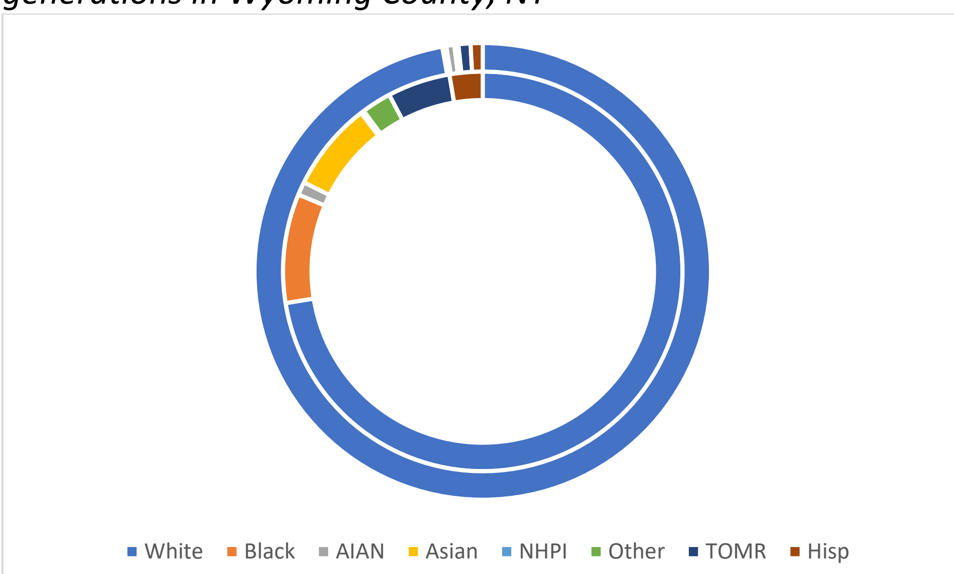
6.5.1 Conclusion

Diversity indices in the demonstration data are higher than in SF1, especially for the rarer category of households with 3 or more generations. This is true for all geographic summary levels considered.

Example: Wyoming County, NY

Outer ring SF1 race of householders of households with 3 or more generations (342 households). Inner ring DHC (271 households)

Figure 25: Race distribution of householders of households with three or more generations in Wyoming County, NY



7 ACCURACY OF THE LARGEST TABLE VALUE

7.1 RESEARCH QUESTIONS:

What can we say about accuracy of the largest value in a table?

In the evaluation of the PL94-171 demonstration data, the Census Bureau set accuracy goals for the largest race/ethnicity groups. This question can be generalized to accuracy of the largest value in a table.

7.2 CONCLUSIONS

The share of the cell with the largest value is on average almost always lower in the demonstration data than in SF1. The larger this maximum value, the smaller the average percentage point difference.

7.3 METHODOLOGY

I used the walkover table to match tables from the demonstration data with SF1 and manually selected as many matches as I could, also based on the number of cells in corresponding tables. No effort was taken to adjust table formats to make even more matches, for example some Group Quarter population by age tables are comparable, but the demonstration data contained fewer cells to limit the numbers of cells with structural zeroes.

Within each table I selected cells that didn't have a further breakdown; no following cells with a larger indent in the walkover table. For example, in the P12 age by sex table the totals by sex are followed by totals by age within this sex that have a bigger indent in the walkover table:

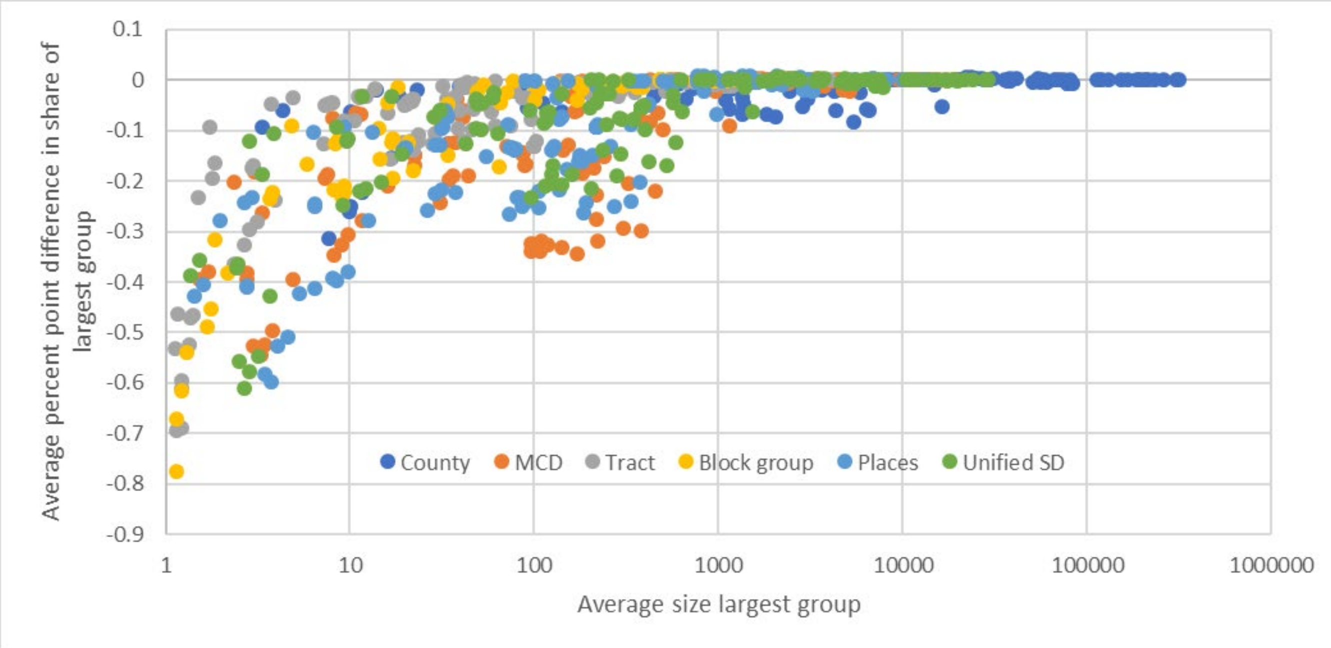
P12				SEX BY AGE FOR SELECTED AGE CATEGORIES [49]
P12				Universe: Total population
P12	P0120001	6	9	Total:
P12	P0120002	6	9	Male:
P12	P0120003	6	9	Under 5 years
P12	P0120004	6	9	5 to 9 years
P12	P0120005	6	9	10 to 14 years

This analyses only used the cells with the biggest indent to find the largest value in the SF1 table. For this cell I calculated a percent error as (DP-SF1)/0.5*(DP+SF1)

For each table and each geography I then calculated an average percent error.

7.4 RESULTS

Figure 26: Average percentage point difference for share of largest table cell



Each dot n the chart represents results from one table and one level of geography. For example the dark blue dot with an X-value of 16229.32 and an Y value of -0.0516 represents County level results for table PCT14B: PRESENCE OF MULTIGENERATIONAL HOUSEHOLDS (BLACK OR AFRICAN AMERICAN ALONE HOUSEHOLDER). The average value of the largest cell in this table is 16,229 and on average the share of this cell is 5.16 percentage point lower in the demonstration data compared to SF1.

Conclusion: The share of the cell with the largest value is on average almost always lower in the demonstration data than in SF1. The larger this maximum value, the smaller the average percentage point difference. It looks like the closer to the spine the smaller the average decrease, but more analyses is needed.

8 DIFFERENCES BY TENURE

8.1 RESEARCH QUESTIONS:

Does the accuracy of the August release of the 2010 Demonstration Data to the original Summary File 1 vary when separated by tenure majority (for example high rental areas compared with high home ownership areas)? Are estimates of households in certain tenure majority areas more vulnerable to inaccuracies than others? Does this depend on the geography of analysis?

8.2 CONCLUSIONS:

- Levels of accuracy varied clearly by tenure, household type, and geography. The most and least accurate tenure type depended on the measure being analyzed and geographic aggregate level.
- Aggregated tracts were consistently more accurate than aggregated block groups for all measures of interest, tenure types, and geographic levels.
- Rental majority areas were highly vulnerable to inaccuracy for households with children and large households (5 or more people)
    - Large households had the highest median absolute percent errors (MdAPEs) of all household types, reaching up to 34.4% in rental majority block groups in Monroe County.
- Errors were significantly smaller for nonfamily households and single person households, but still displayed differential patterns of accuracy.
    - The largest MdAPEs for these household types occurred in owner majority areas across all geographies, except for single person households in New York state where the least accurate tenure type was rental majority.

8.3 METHODOLOGY:

- We used the 2010 Summary File 1 and the August 2022 release of the 2010 Differential Privacy DHC housing unit files at the Census Tract and Block Group levels of geography, excluding Puerto Rico.
- This analysis was limited to census tracts with 200 or more households and block groups with 150 or more households to exclude special use areas.
- Majority housing tenure was determined by calculating percent ownership in a tract or block group: [(IFF002_sf + IFF003_sf)/H8C001_sf] *100
- Geographies were classified into the following categories based on the share of owner-occupied households: rental majority if the share of ownership was <=20%, mixed tenure if the share of owned households was between 21% and 79%, and majority owned if >=80% of occupied households were owned.
- Our final dataset contained 71,842 Census tracts and 214,558 block groups.

8.3.1 Metrics of error:

This section of analysis focuses on one key metric, Median Absolute Percent Error (MdAPE) as a measurement of accuracy to the original Summary 1 File. We utilize the median rather than the mean to be robust to outliers. In our calculations of error, cases where the original SF1 was 0 and the DHC was nonzero resulted in an observation of 0.5.

8.4 Results

8.4.1 Comparing the Accuracy of the 2010 SF1 to DHC: Households with Children (<18)

Chart 1: MdAPE of Census Tracts, by Aggregate Level

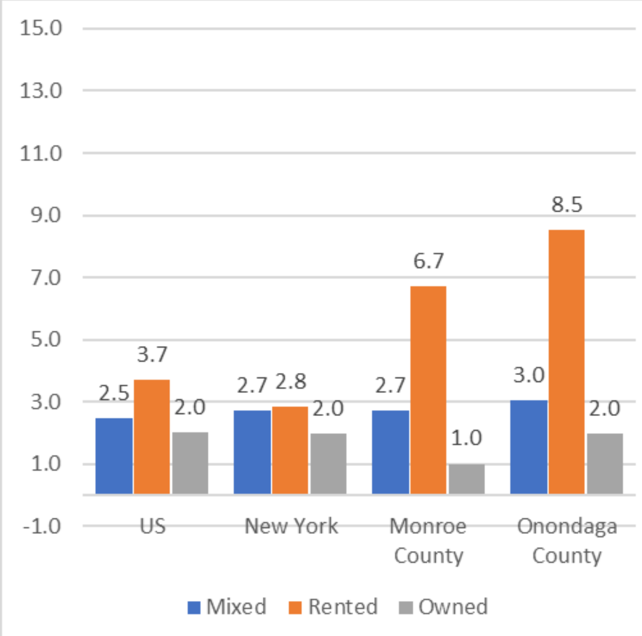
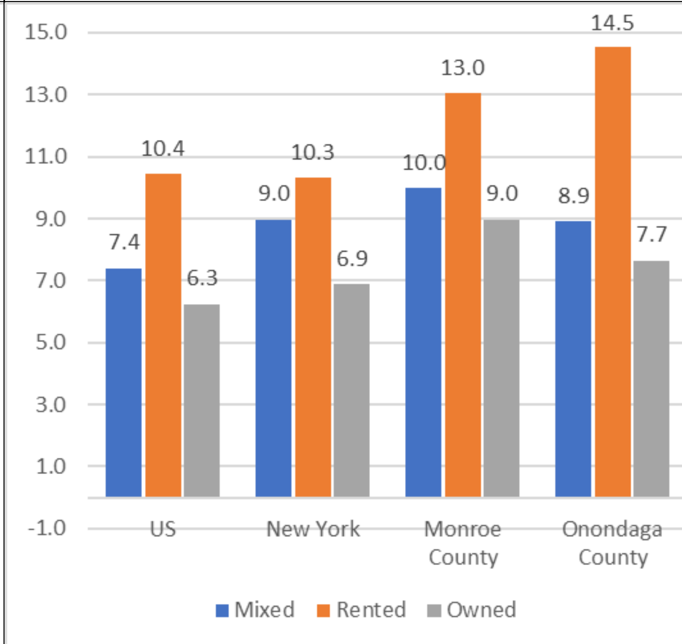


Chart 2: MdAPE of Block Groups, by Aggregate Level



Conclusions for Charts 1 & 2:

-Block group-level counts of households with children under 18 were less accurate to the original SF1 than tract-level counts for all tenure categories and aggregate levels.

- At the tract level, the median absolute percent error (MdAPE) between the original 2010 SF1 and the 2010 DHC for households with children in rental majority areas ranged from 3.7% to 8.5%, while this range at the block group level was from 10.4% to 14.5%.

- Among census tracts, the MdAPE for owner majority and mixed tenure areas were low, ranging from around 1% to 3%, but still rose to between 6% and 10% at the block group level.

- At both the tract and block group level, counts of households with children under 18 in rental majority areas were less accurate to the original SF1 than owner majority and mixed tenure areas.

Estimates of households with children in owner-majority areas were consistently the most accurate at every aggregate level (i.e. national, state, and county).

8.4.2 Comparing the Accuracy of the 2010 SF1 to DHC: Non-family households

Chart 1: MdAPE of Census Tracts, by Aggregate Level

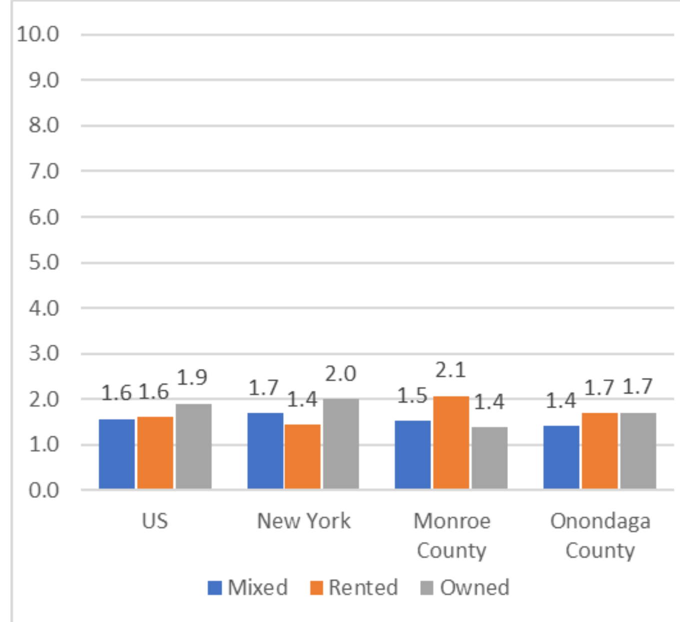
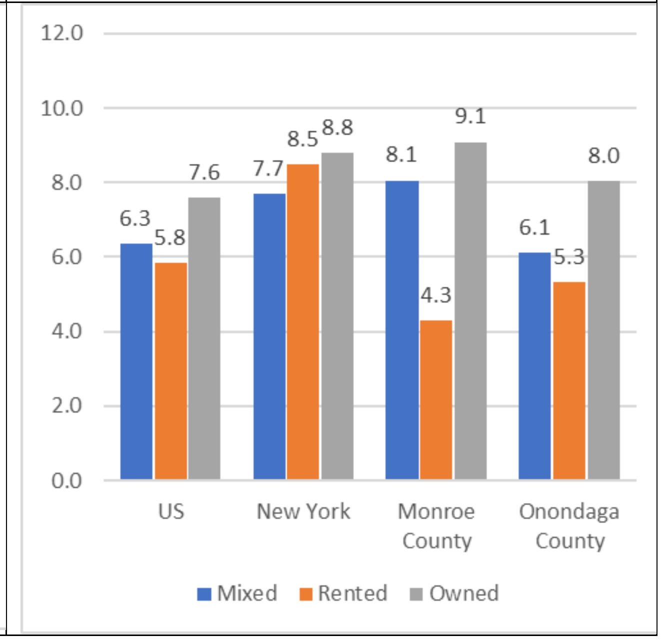


Chart 2: MdAPE of Block Groups, by Aggregate Level



Conclusions:

-Though block groups remained less accurate than tracts, patterns among non-family households by tenure and geographic level differed from those observed for large households and households with children.

- At the tract level, differences between the files remained low. However, owner majority areas had the highest MdAPE for all geographies except for Monroe County, where the largest MdAPE was found among nonfamily households in rental majority areas.

- Owner majority areas at the block group level were the least accurate for all geographies, ranging from 7.6% in the U.S. to 9.1% in Monroe County. For block groups, households in rental majority areas of Monroe County had the lowest MdAPE.

- For nonfamily households, block groups aggregated to the New York State level were less accurate at all tenure areas than nationally aggregated block groups. Conversely, Onondaga County block groups were more accurate than New York State for all tenure types.

8.4.3 Comparing the Accuracy of the 2010 SF1 to DHC: Single Person Households

Chart 1: MdAPE of Census Tracts, by Aggregate Level

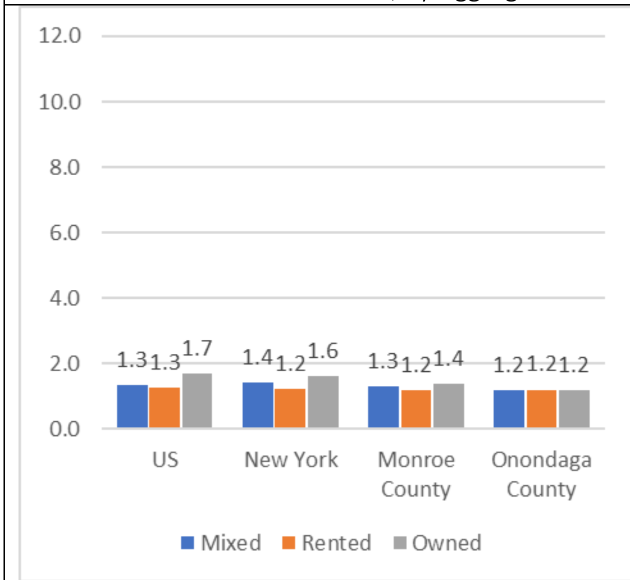
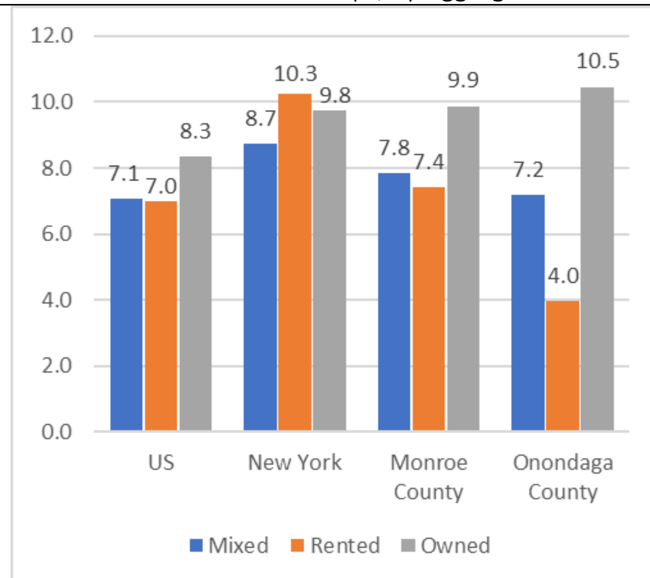


Chart 2: MdAPE of Block Groups, by Aggregate Level



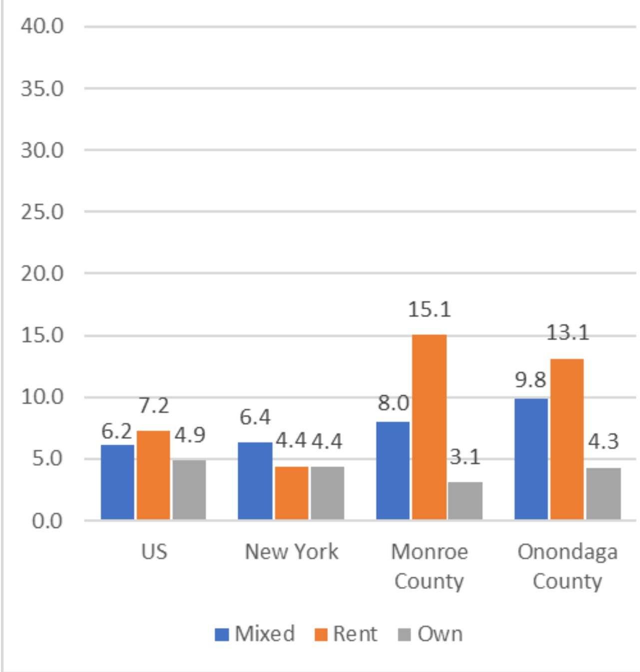
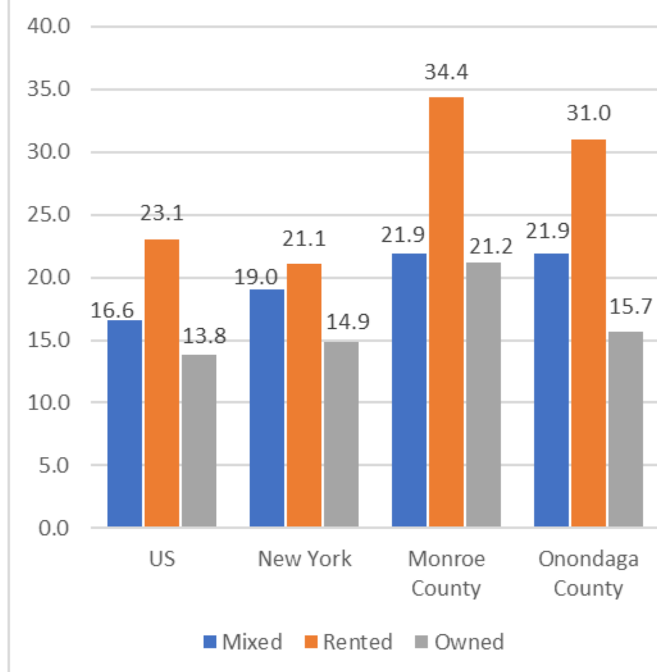
-Patterns of accuracy for single person households mirrored those of nonfamily households.

- Errors among aggregated tracts were consistently low.
- Single person households in owner majority block groups were the least accurate for most geographies, ranging from 8.3% in the U.S. to 10.5% in Onondaga County.
    - State level aggregated block groups were the exception, with rental majority areas producing the highest MdAPE (10.3%). For all other geographies, single person households in rental majority areas were the most accurate.

As with nonfamily households, block group estimates of single person households in New York State were less accurate at all tenure areas than at the national level.

- Both Onondaga and Monroe County had lower MdAPEs than New York State for single person households in mixed and rental majority areas.
- MdAPEs of single person households in owner majority areas were greater at the county level than the state level.

8.4.4 Comparing the Accuracy of the 2010 SF1 to DHC: Large Households (5+ People)

Chart 1: MdAPE of Census Tracts, by Aggregate Level	Chart 2: MdAPE of Block Groups, by Aggregate Level
	

Conclusions:

-Of the four variables examined in this analysis, estimates of large households in the DHC were the least accurate to the original SF1.

-At the tract level, the median absolute percent error (MdAPE) between the original 2010 SF1 and the 2010 DHC for large households in rental majority areas ranged from 4.4% to 15.1%, while this range at the block group level was from 21.1% to 34.4%.

- Among census tracts, the MdAPE for owner majority areas ranged from 3.1% to 4.9%, but still rose to between 13.8% and 21.2% at the block group level.

- Similarly, counts of large households in mixed tenure areas had MdAPEs between 6.2% and 9.8% at the tract level and between 16.6% and 21.9% at the block group level.

- Though large households in rental majority areas had the highest MdAPE across all geographies, this was not the case at the tract level.

- For large households in tracts aggregated to New York State, mixed tenure areas were the least accurate while renter and owner majority areas had the same MdAPE.