STATE OF WASHINGTON

OFFICE OF FINANCIAL MANAGEMENT

Insurance Building, PO Box 43113 • Olympia, Washington 98504-3113 • (360) 902-0555

May 27, 2021

Disclosure Avoidance System Team:

Thank you for continuing to provide us with the tools we need to evaluate the effects of differential privacy (DP) on the usability of the 2020 Census products. We wrote an earlier letter on December 21, 2020, based on the information we found in the November 16, 2020, Privacy Protected Microdata File. This letter revisits the concerns we found in that document in light of the latest PPMF files released April 28, 2021. The increase of the privacy budget is greatly appreciated and of course the PPMF file using that budget mitigates many issues. As you intended, the release of the PPMF using the original budget helps us evaluate whether processing biases have been reduced.

This PPMF release is still limited to the P.L. 94-171 redistricting data product. We are still extremely concerned that decisions about the privacy loss budget are being made without giving stakeholders the opportunity to provide feedback on the detailed data that includes 5-year age, sex, race and ethnicity data, household roster, and other person/household join tables. The amount of concern has increased now that we know that subsequent products will be controlled to the P.L. 94-171 data.

Illogical and implausible values remain in the most recent PPMF data release. It is clear that substantial effort was made in addressing implausible values. None were eliminated but many were reduced in size even when using the original epsilon. However, there was an increase in the population assigned to blocks that had no occupied housing units and those that had no population in the published 2010 data. The 2020 census data will simply not be credible if illogical and implausible values exist, even if their frequency is reduced.

We still believe that the correct approach is making occupancy status invariant, limiting the distribution of household population to blocks with occupied housing units, forcing each occupied unit to have at least one person, and limiting the frequency of outlier housing unit sizes. These changes would go a long way towards improving the validity of the data.

City occupancy and persons per housing unit rates have been substantially improved to the point that they are now acceptable for one, but not all, of our use cases.

Accuracy of the racial diversity has improved but there are still concerns particularly in the multi-race categories.

We believe the root cause of many of these issues is the independent processing of the housing and population data that DP currently employs. We urge you to re-think that choice. Processing the data in this way removes all household relationships from the data tables. Those relationships are extremely important and are the source of very valuable information for a large number of census data products.

We believe the de-coupling of this relationship will have far-reaching consequences affecting not only the decennial census data but also the American Community Survey data.

Illogical Values and Related Biases

The illogical distribution of occupied housing units continues to create errors that must be corrected if the data are to be viewed as credible. The new DP process has actually increased the cases of people living in blocks with no occupied housing units. Using the latest version with the original epsilon, the number of these cases in Washington increased from 9,197 blocks in the November release with 174,071 people to 12,553 blocks with 268,395 people. With the higher epsilon version, it increased less to 10,949 blocks with the population affected reduced to 148,976. The inverse situation was dramatically improved however. With the higher epsilon there are only 730 blocks with occupied housing units and no population. This is down from 8,978 in the November release. This appears to be mostly due to processing improvements as most of the improvement is seen with the lower epsilon.

The new DP process made no improvements for housing unit size at the block level. In the November release there were 25,005 people in 500 blocks with an implausible housing unit size greater than 20 people. Without the benefit of the higher epsilon, that number increased to 32,371 people in 630 blocks. With the higher epsilon there is a slight improvement to 19,451 people in 472 blocks. We reiterate that any of these illogical and improbable situations limit the analysis of this data as people and occupied housing involved would have to be reassigned to correct these issues. For the present iteration of the PPMF, we find these inconsistencies very troubling.

In Washington, the bias in the occupancy rates at the census block level was not improved at all. Interestingly, the lower epsilon version did improve occupancy rates slightly from an average 0.042 persons per housing unit lower than the published 2010 data to 0.039 lower. The higher epsilon version has exactly the same value as the November version. Occupancy rates range from 0 to 1 so this is a 4.2% bias towards lower occupancy. The standard deviation of the occupancy rate did not improve significantly.

To calculate the housing unit size (PPH), the population in housing units is divided by the number of occupied housing units. The accuracy of these values worsened in this new version. *At the block level, the average PPH was 0.482 higher than it was in the original 2010 data.* In the new version with the higher epsilon it is 0.516 higher. The new process is actually worse. In the lower epsilon version, it was 0.698 higher than the published 2010 census data. This means there is nearly half a person more per housing unit. As with the occupancy rate, if the effects of the DP process were random one would expect this value to be near zero. The standard deviation for housing unit size did improve from 3.333 to 2.675 in the higher epsilon version but with the same epsilon it went up to 3.569. Even at 2.675, this is over three times more variance than the published 2010 value of 0.840.

The block level noise from the reassignment of vacancy status combined with the decoupling of housing unit and population estimates has far reaching impacts. *This is of great concern because we use these rates for our small area estimates (below the county and city levels).*

There has been a substantial improvement of occupancy and PPH rates at higher levels of geography. The bias at the city level has been eliminated. The average of the differences in the occupancy rates and PPH from the 2010 published rates are on average zero. The standard deviations of those rates are also nearly identical. This does not mean the rates have not been affected. However, the rates are dramatically improved. The average of the absolute difference in the occupancy rate dropped from 0.051 in November to 0.012 with the higher epsilon. The same measure for PPH dropped from 0.262 to 0.041.

We reran our estimates to understand how different our 2020 city housing unit method estimates would have been if the error in the latest PPMF had been present in the 2010 Census. For cities under 1,000 people, the difference dropped from an average of 15.6% to an average of 1.7%. The worst city was off by 10.5% with the higher epsilon where in the November version one city had a population difference of 93.5%.

We would like to thank your team for their work in addressing city-level housing unit occupancy and PPH rates as it is one our critical business needs.

In the November version there are 8,436 blocks in Washington where the entire population of 94,444 persons was under the age of 18. Changes in the DP process brought that number down to 7,446 children in 2,217 blocks. After the increase in epsilon that went down to 2,278 children in 1,112 blocks. This is a vast improvement. However, it is still an anomaly that will reduce confidence in the 2020 numbers as these living situations are extremely rare in reality. In fact, there only 11 such blocks with 186 people in the original 2010 Census data.

Race and Ethnicity

The apparent flaw in the DP process that was systematically decreasing racial diversity has been reversed and it is now increasing diversity by similar percentages. We used a simple metric for change in racial diversity. It simply counts the number of race alone categories were represented in a geography. In the November release, 31% of Washington blocks saw a reduction in the number of races represented, while only 6% saw an increase. Those numbers have nearly inverted with the high epsilon version of PPMF with 11% of blocks showing fewer races and 31% showing more races. This reversal of the bias weakened but continued up to the city level with 12% of cities increasing and 16% decreasing. At the county level the bias goes away with 3% increasing and 3% decreasing. While reduced in effect, this is still a serious issue for anyone trying to use census data to understand racial representation and diversity.

The correlation between the percent of people in a city that are of Hispanic origin and the average PPH of that city improved from and R-squared of 0.394 to one of 0.497 in the high epsilon version. This brings it closer to the original 2010 Census value of 0.600. All the improvement was solely due to the increased epsilon as the low epsilon version had a slightly reduced R-squared of 0.346. The change in the power of the correlation was a little weaker. In the published 2010 Census data, the power of the correlation was 1.87. The November PPMF had a power of 2.03 and the high epsilon version a power of 2.05.

There was no real improvement at the block level going from an R-squared in November of 0.002 to 0.006 in the high epsilon version. The published 2010 census data had an R-squared value of 0.120.

This statistical correlation is just representative of many others that should be strong in the data but have been stripped by the new DA process. There have been many such statistical correlations used in numerous studies and estimation processes over the past three decades. By focusing on a handful of on-going use cases you will be missing the many small and/or one-time studies which will no longer be possible due to these changes.

Rural – Urban bias

In our feedback letters for the November 2019 and May 2020 demonstration data we expressed concern about a bias that was causing blocks with small populations to get larger and blocks with large populations to get smaller. A result of that bias was areas comprised of lots of blocks with small populations, i.e. rural areas, were increasing in population size while urban areas were decreasing. This bias was much improved in the November PPMF over the previous releases. It had 1,824 people being shifted from urban areas to rural area. However, this bias has increased in the April PPMF. The lower epsilon PPMF shows an exchange of 11,834 people and the higher epsilon PPMF shows 6,968. *Both April files appear to be a step backwards with respect to this bias.*

Redistricting

After the November version of the PPMF was released much attention was given to the redistricting use case. Unfortunately, this has not resulted in improved results for Washington's 49 legislative districts. It has in fact led to an increase the absolute difference in population from the census results by 669%. In the November version the absolute average difference between 2010 Census published results and the November PPMF was 39 people. In the new high epsilon version it was 260. While the average percent change is only 0.19, redistricting is about counting whole people to insure equitable representation. Numbers this far off would result in real changes to choices made in the redistricting process.

We looked at the percent minority population by legislative district and were again pleased to find that these values were not changed from the original 2010 published results in a way that would likely affect the redistricting process. The maximum change in percent minority went down from 0.81% to 0.50%.

Voting precincts are much smaller units than legislative districts. Washington has 6,696 voter precincts. These precincts are a key component of all voting jurisdictions in the state. In the higher epsilon PPMF, the number of precincts that would have been flagged as not having voting age population, when they in fact did, dropped from 22 precincts to 3. The inverse value stayed the same with the PPMF data having 2 precincts with voting population that had none in the published 2010 data.

The racial makeup of voting precincts also changes. The reversal of the bias towards the reduction of racial diversity noted in an earlier section also occurred at the precinct level. In the November release there were 4,512 precincts with fewer races represented, while 318 showed an increase in the number of races. In the high epsilon PPMF there is a weaker bias with 1,047 precincts with fewer races represented and 1,670 with more. There were fewer precincts switching from majority white non-Hispanic to a majority non-white or Hispanic and vice-versa. There were 136 precincts switching from majority white non-Hispanic to majority non-white or Hispanic in the November data, but only 29 in the new high epsilon file. There were 73 precincts that went the other direction from majority non-white or Hispanic to majority white non-Hispanic in the older data. That went down to 47 in the new high epsilon version. This represents an improvement in usability for redistricting, but still poses some problems.

Group Quarters

The error introduced in census blocks with group quarters has been reduced, especially the demographic detail. In our December letter we concentrated on the effect that changes in population in blocks with group quarters would have on redistricting. The improvement of the demographic detail in this release in particular will improve its usability.

Our office also has a critical use case involving group quarters. As part of our postcensal city and small area estimation processes we track the population of many, but not all, of Washington's group quarters. In order to use these estimates we must identify the census block in which each facility is located and subtract our known count from the Census' group quarter count for that block. This way we can retain the information for any of the untracked group quarters that might also be in that block. In past censuses this process was challenging as we frequently had the facility in a different block than the census and had to make adjustment accordingly. Now that the group quarters population is not held invariant, it will be difficult to determine which blocks contain the small group quarters we track. It will also be difficult to determine if a block has only our known group quarters or if the Census counted another facility we aren't tracking.

We identified 138 blocks in Washington that contained major group quarters in 2010. In those blocks the average absolute percent error of the total population, as compared to the published 2010 Census, dropped from 5.5% in the November release to 3.6% in the high epsilon version. The average absolute percent error of the age by race detail cell dropped from 39.1% to 15.9%.

This improvement will help with our group quarters processing, but difficulties will remain. The following tables show the selection of blocks with major correctional facilities in Washington. The last four records in each table are juvenile facilities and the rest are adult. The most noticeable change between them is the large reduction in the error in the black population.

	2010 PL				2010 November PPMF				Error Introduced by DP			
Block	Total Pop	GQ Pop	Under 18	Black	Total Pop	GQ Pop	Under 18	Black	Total Pop	GQ Pop	Under 18	Black
530090002001051	896	896	0	244	911	911	3	252	15	15	3	8
530110405041069	219	219	0	39	223	223	1	35	4	4	1	-4
530210208003025	1445	1445	0	266	1457	1457	0	266	12	12	0	0
530270016002013	2005	1967	5	372	2004	2004	0	340	-1	37	-5	-32
530319507021027	377	377	0	72	335	335	0	0	-42	-42	0	-72
530459604001012	230	178	17	20	233	142	0	2	3	-36	-17	-18
530459606001008	1673	1673	3	315	1671	1671	1	293	-2	-2	-2	-22
530530725042008	827	827	0	144	838	838	0	143	11	11	0	-1
530610522091002	2466	2463	2	489	2461	2459	14	466	-5	-4	12	-23
530630104011011	2184	2184	0	320	2216	2216	0	309	32	32	0	-11
530719204001001	2314	2300	4	552	2312	2264	0	594	-2	-36	-4	42
530330326023045	126	126	120	25	90	90	89	24	-36	-36	-31	-1
530419710001027	193	193	103	60	168	168	90	0	-25	-25	-13	-60
530499504002046	156	98	94	17	146	54	74	0	-10	-44	-20	-17
530670127205045	203	203	156	35	181	181	167	0	-22	-22	11	-35

	2010 PL				2010 April 12.2 PPMF				Error Introduced by DP			
Block	Total Pop	GQ Pop	Under 18	Black	Total Pop	GQ Pop	Under 18	Black	Total Pop	GQ Pop	Under 18	Black
530090002001051	896	896	0	244	878	878	0	243	-18	-18	0	-1
530110405041069	219	219	0	39	226	226	3	38	7	7	3	-1
530210208003025	1445	1445	0	266	1411	1411	1	259	-34	-34	1	-7
530270016002013	2005	1967	5	372	2002	1924	26	362	-3	-43	21	-10
530319507021027	377	377	0	72	373	373	0	65	-4	-4	0	-7
530459604001012	230	178	17	20	231	182	14	23	1	4	-3	3
530459606001008	1673	1673	3	315	1653	1653	0	310	-20	-20	-3	-5
530530725042008	827	827	0	144	805	805	2	149	-22	-22	2	5
530610522091002	2466	2463	2	489	2465	2438	16	498	-1	-25	14	9
530630104011011	2184	2184	0	320	2168	2168	0	321	-16	-16	0	1
530719204001001	2314	2300	4	552	2333	2274	23	557	19	-26	19	5
530330326023045	126	126	120	25	94	94	94	27	-32	-32	-26	2
530419710001027	193	193	103	60	153	153	85	52	-40	-40	-18	-8
530499504002046	156	98	94	17	153	76	88	14	-3	-22	-6	-3
530670127205045	203	203	156	35	146	146	129	26	-57	-57	-27	-9

Conclusion

Thank you again for the opportunity to review this data and provide feedback. We appreciate the lengths that the Census Bureau DAS team has gone to protect privacy and the work you have put in.

We believe the independent processing of population characteristics and housing characteristic is having a negative impact on the PPMF data. The relationship between population and housing is a fundamental to demographic modeling. We believe that population and housing relationships should be preserved.

We believe that illogical and implausible occupancy rates and household sizes in the block data will result in credibility issues and prevent users of block data from performing their work. As we stated, making occupancy status invariant and/or eliminating illogical/improbable values during the post processing would go a long way to improving the usability of the block data.

We appreciate the improvements to city level occupancy and household size, as well as improvements to group quarter values. Work is still needed to help preserve the distribution of race and ethnicity characteristics at sub-county geographies as well as reducing the bias that is causing rural areas to increase in population size.

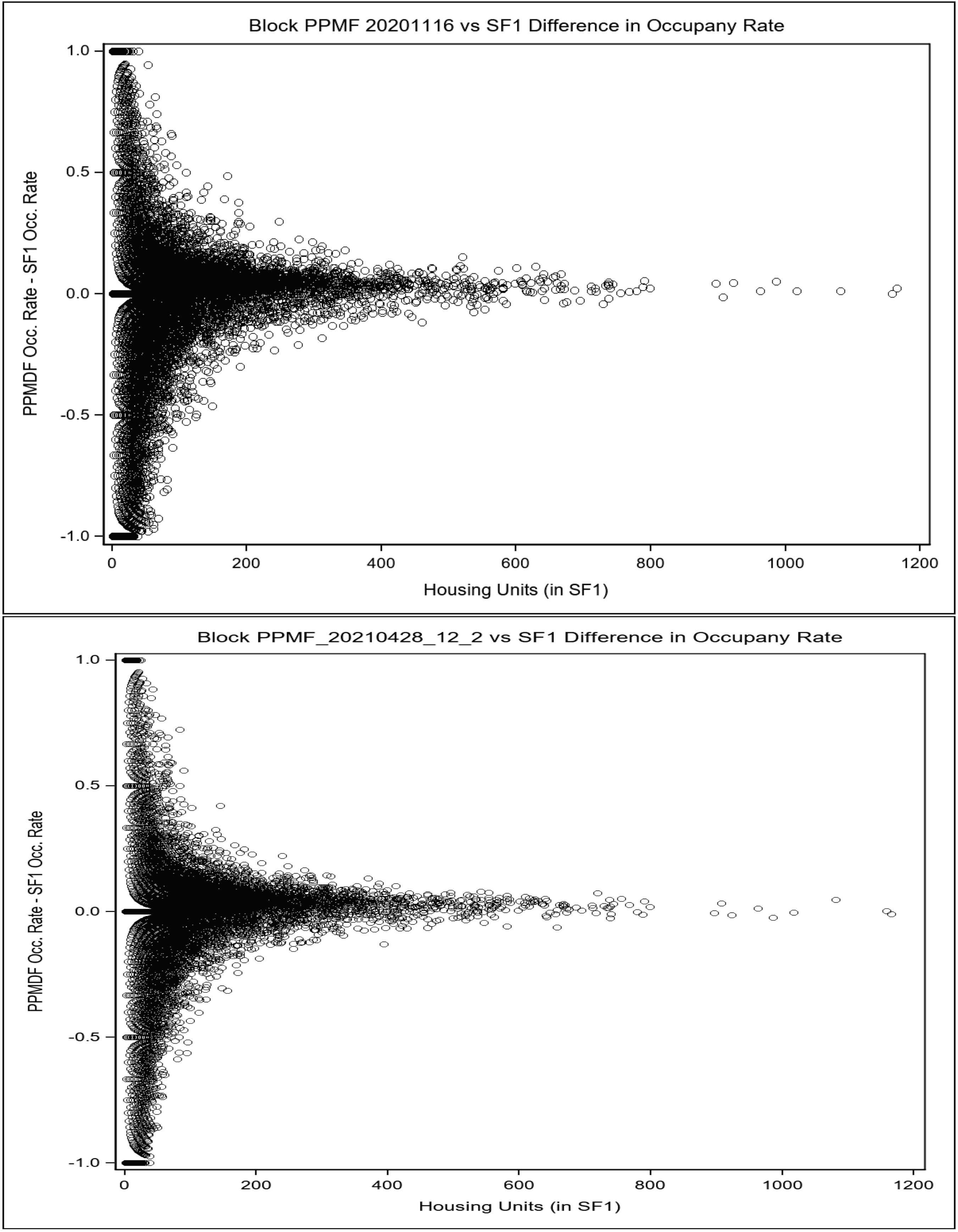
Sincerely,

Mike Mohrman, State Demographer
Forecasting and Research Division

cc Marc Baldwin, Assistant Director

Appendix

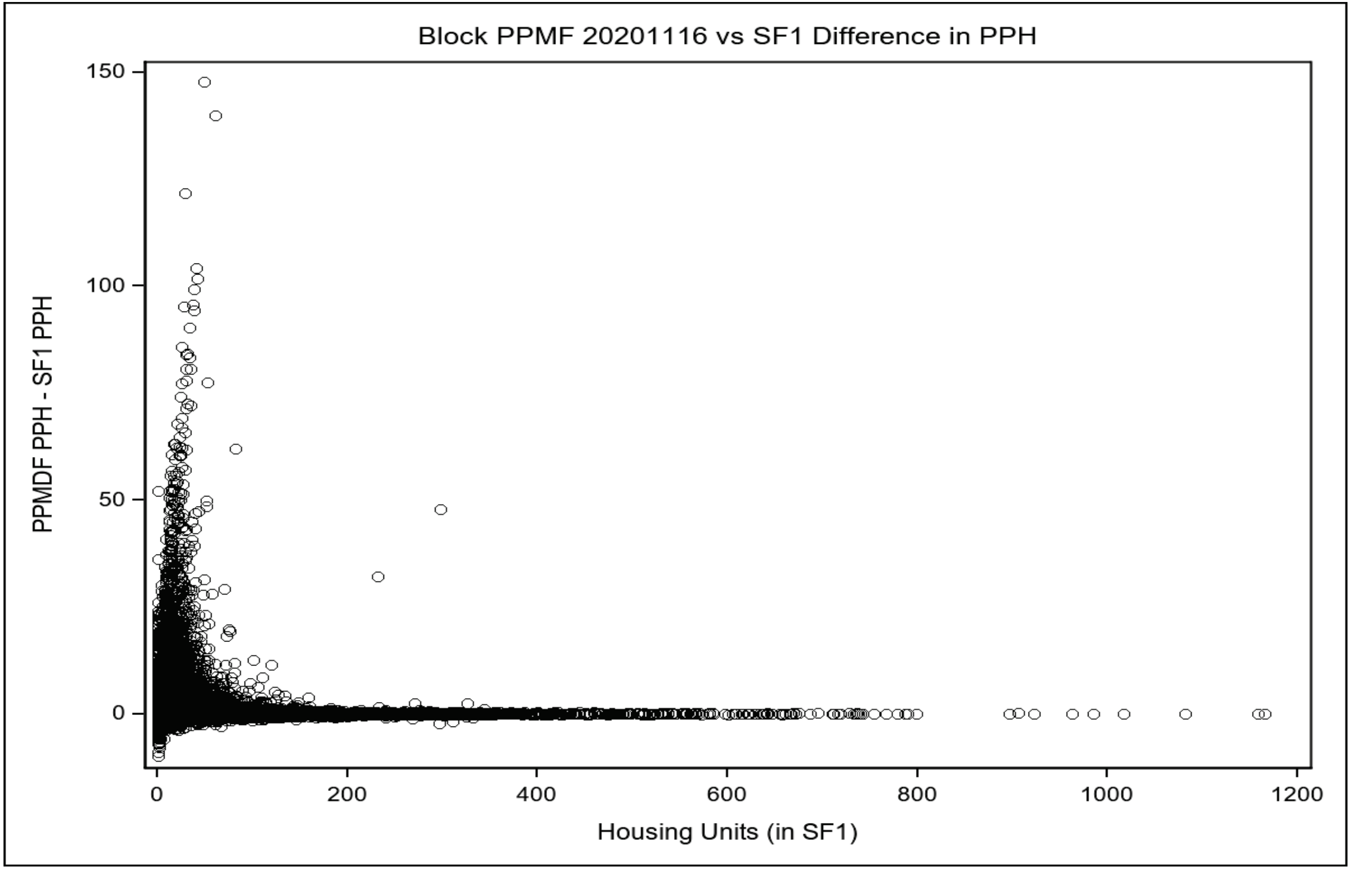
Figure 1



Block PPMF 20201116 vs SF1 Difference in Occupany Rate

Block PPMF_20210428_12_2 vs SF1 Difference in Occupany Rate

Figure 2


Block PPMF 20201116 vs SF1 Difference in PPH
PPMDF PPH - SF1 PPH
150
100
50
0
0
200
400
600
800
1000
1200
Housing Units (in SF1)

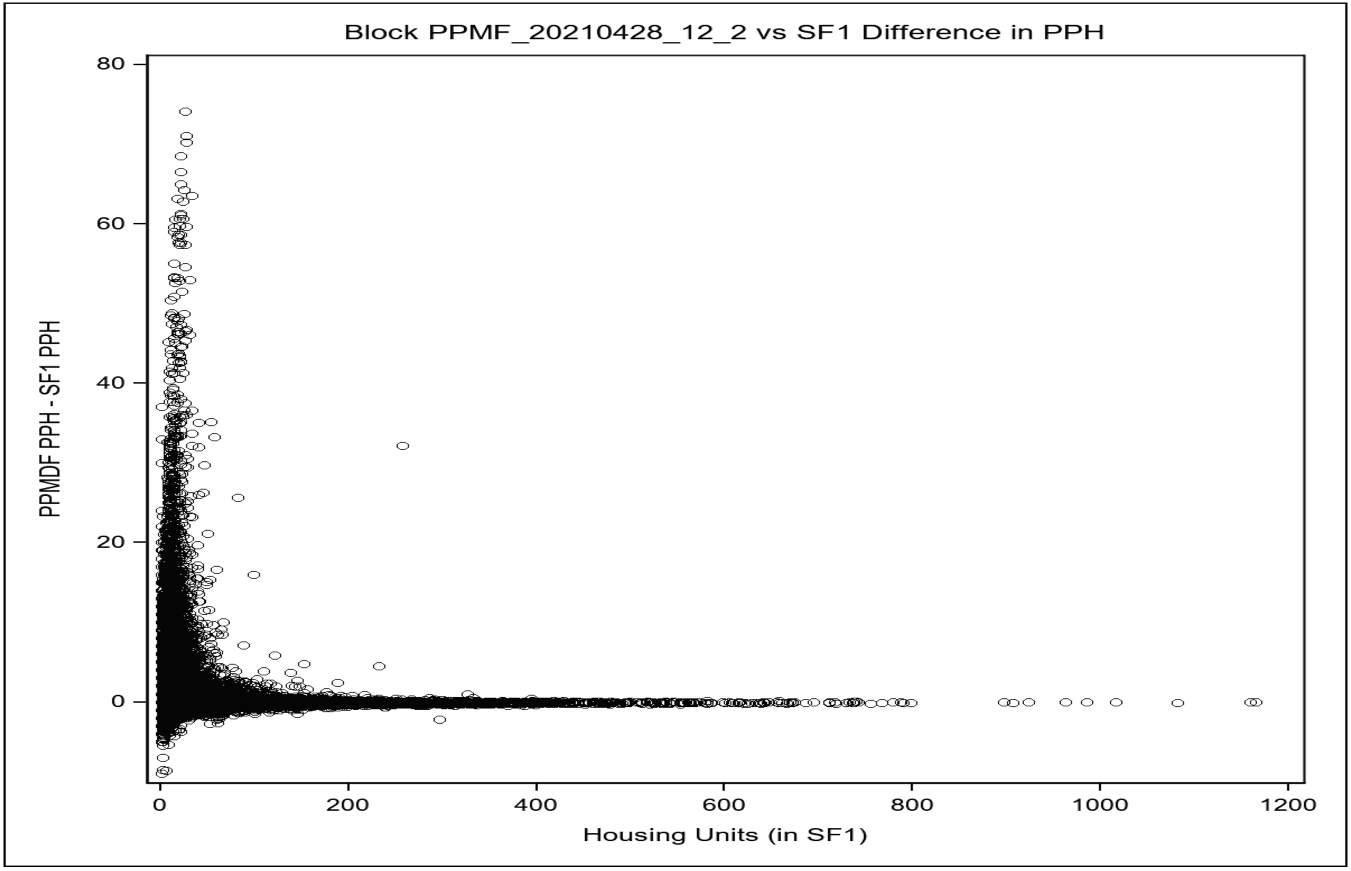
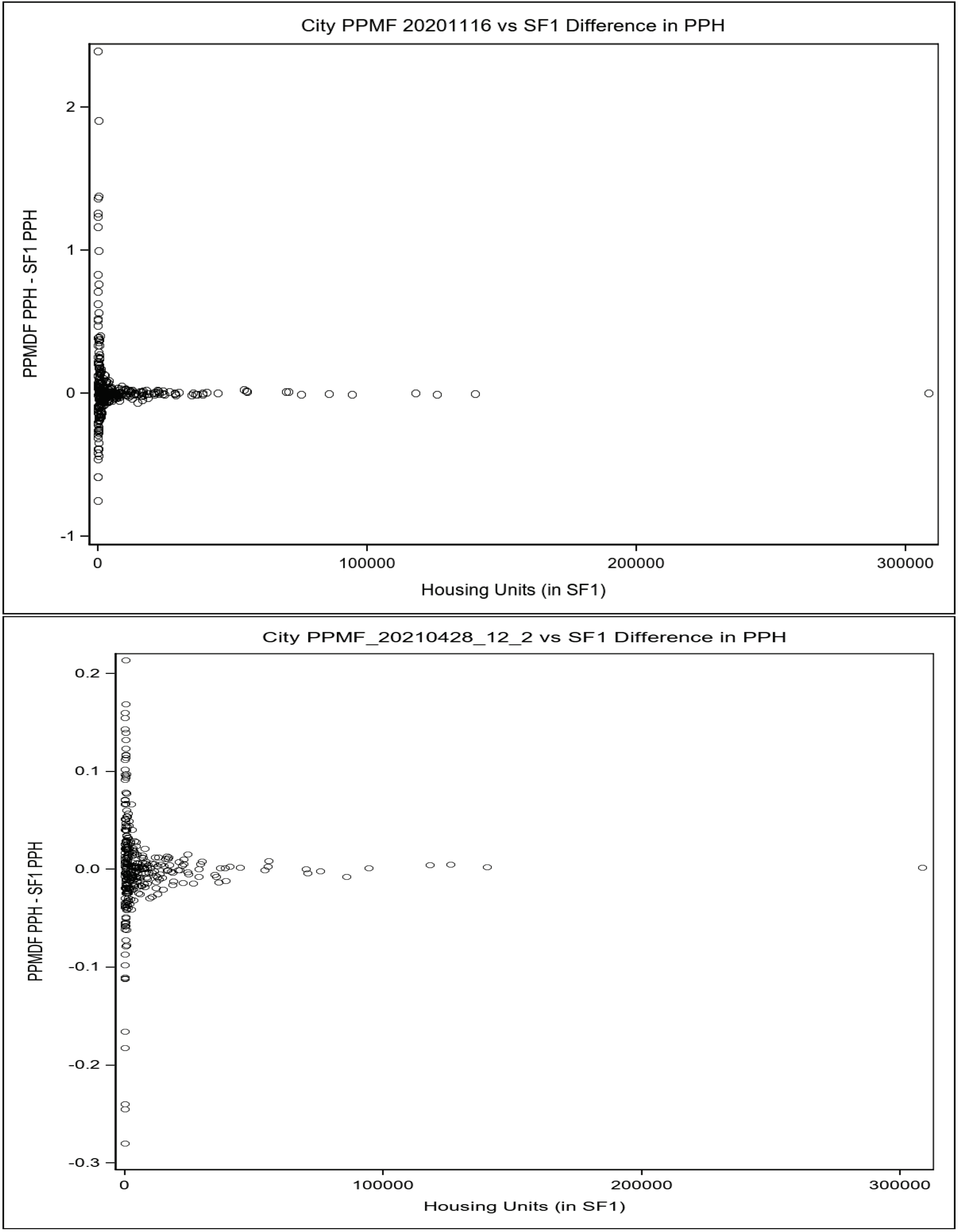

Block PPMF_20210428_12_2 vs SF1 Difference in PPH
PPMDF PPH - SF1 PPH
80
60
40
20
0
0
200
400
600
800
1000
1200
Housing Units (in SF1)

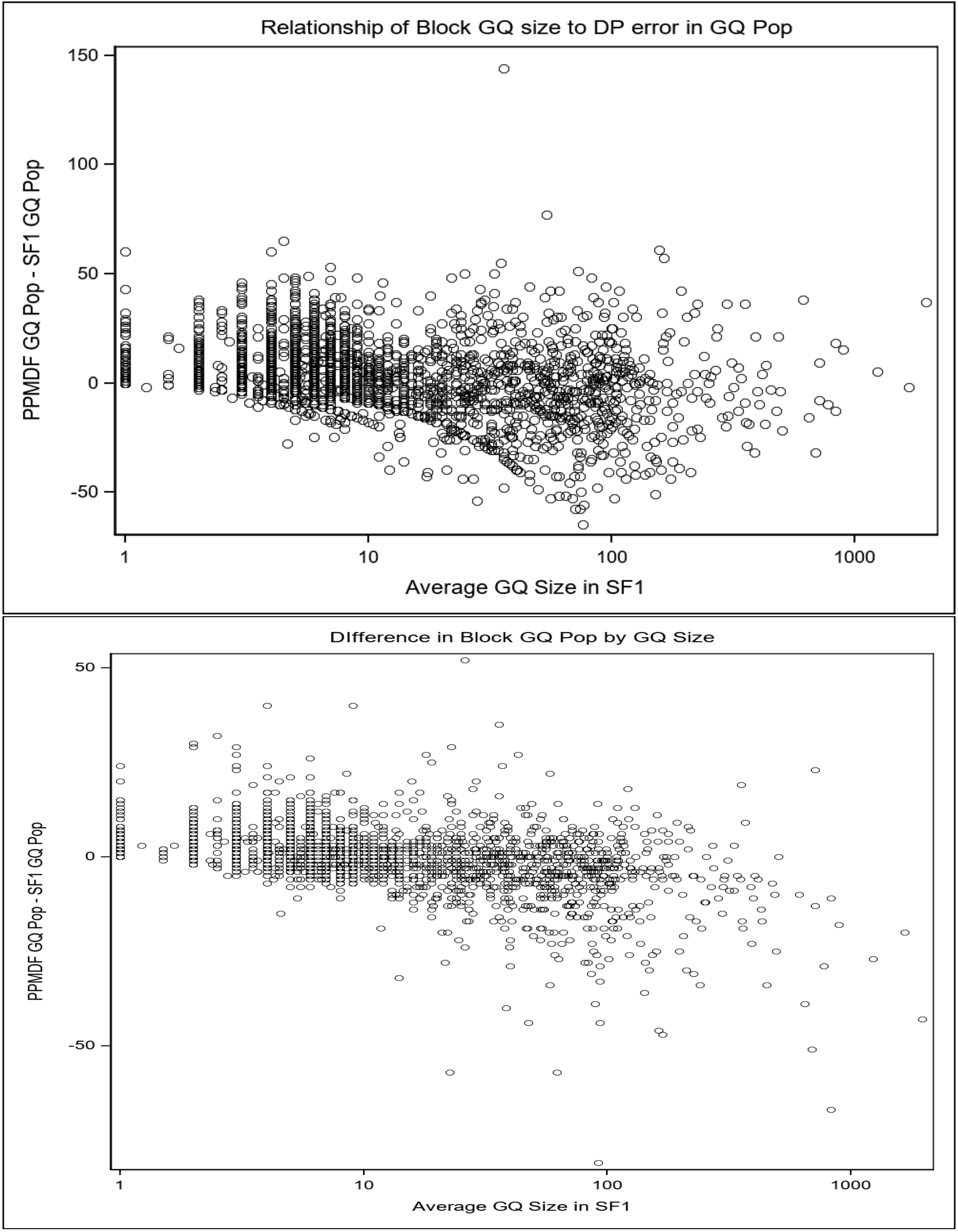
Figure 3



City PPMF 20201116 vs SF1 Difference in PPH

City PPMF_20210428_12_2 vs SF1 Difference in PPH

Figure 4


Relationship of Block GQ size to DP error in GQ Pop
PPMDF GQ Pop - SF1 GQ Pop
150
100
50
0
-50
1
10
100
1000
Average GQ Size in SF1
DIfference in Block GQ Pop by GQ Size
PPMDF GQ Pop - SF1 GQ Pop
50
0
-50
1
10
100
1000
Average GQ Size in SF1