2020 CENSUS: PRIVACY CONCERNS

Over the past year or so, we have been giving a heads up that the 2020 census data was to be affected by new and more stringent non-disclosure rules intended to protect the identify of census respondents. We recognize that such protection is important, but the effects will be felt for years to come. This report explores the root of the problem, and shows detail at the block level.

AGS
APPLIED GEOGRAPHIC SOLUTIONS
Powering Smarter Market Decisions™

www.appliedgeographic.com

The Problem

In years past, the census has – as required by law – made substantial efforts at protecting the privacy of individuals. As the genealogy world well knows, the physical records which have names and addresses, are sealed for decades. When the census included both the short form and the long form, the sensitive personal data found in the long form was reasonably well protected -- it was based on a sample and techniques were employed to “borrow” characteristics between similar, nearby census blocks. With the demise of the long form – replaced by the American Community Survey (ACS) – the census consists of only completely (obviously with some error) enumerated geographic areas. As a result, the data for small areas can be used in conjunction with other databases (mailing lists, property records, etc.) to potentially identify individuals within them.

The unpleasant conclusion is that the data has been seriously corrupted, so much so that a significant number of census block groups have statistically impossible data, among them:
·entire blocks of unsupervised children in households (no adults)
·ghost communes, where there are occupied dwellings with no people
·families well above average household sizes

For every identified impossibility, there lurks underneath it at least ten improbabilities, and this is just the baseline numbers. The real meat of the 2020 census is found in the detailed tables which address key population characteristics (age, sex, race, Hispanic origin, ancestry) and household characteristics (household size and structure).

The privacy “budget” was essentially exhausted at the block group level with the release of the general population counts, and the Census is considering releasing the detailed tables only to the Census Tract level. It is not hard to understand why. Massive reallocation was required just to release top level statistics. Imagine what will need to be done to publish a table of population by age and sex?

In such a table, the worst-case scenario is a value of one: showing that there is one female age 20-24 in a block allows that individual to be identified. At a certain level of geographic aggregation, the data must match the actual totals – the offending cell must be modified, and this means that the value must be changed in the opposite direction for some nearby block. Even at the block group level, there will be a great many cells with a value of 1 or 2, and each adjustment affects at least two block groups. Multiply this through, and you quickly see how pervasive the issue becomes.

From an operational standpoint, the goal of maintaining privacy while maintaining the essence of each geographic unit is an almost impossible task. The published redistricting results clearly indicate that the problem was not solved by one at a time characteristic trading between nearby areas but instead relied upon bulk operations which radically change the essential character of each geographic unit. The presence of statistical impossibilities is clear evidence of this.

Coming Issues with the ACS

Logic would suggest that the ACS analysts would have access to the original census data to both structure their sampling and extrapolate the results. Given some of the comments and discussions we have seen, this is not necessarily the case.

Alas, all is not well in ACS-land. The 2020 1-year series is delayed until the end of November and is being touted as “experimental”. We expect that the tables will be little more than asterisks punctuated by the occasional numeric entry. The 2020 survey occurred at the pandemic peak and response rates were substantially lower than usual. Worse, in-person visits were cancelled, and the results are seriously biased. The 2021 ACS should be much improved, but likely not at normal quality levels.

Don’t expect the 1-year series to be back to normal until late 2023, and the critical 5-year series until 2026.

What remains is a census made far less usable by privacy concerns, and an ACS series with noticeable deficiencies for the next several years.

Our philosophy at AGS has always been to educate users about both the strengths and weaknesses of all databases we create or provide, and we do not shy away from the concept of error. Error is simply uncertainty, and no data is without error. Higher error rates do not render a dataset unusable; they simply increase our level of uncertainty about decisions we may make using it.

The ACS will recover from its pandemic issues, but this will take a few years to work through. In the meantime, our goals are as follows:

- to educate both our business partners and end users on the nature and scope of the issue and, equally importantly, its impact on decision making processes
- to utilize the geostatistical techniques that we have developed over the decades to enhance the usability of the data as much as possible by reigning in the statistical impossibilities and using multi-tiered maximum likelihood models to provide a consistent and reasonably accurate benchmark point

Here at Applied Geographic Solutions, we have been working with census data for several decades and have developed and refined a powerful set of tools for analyzing and manipulating small area data.

Spatially Aware Matrix Mathematics

Users of census data have long faced the issue of the ever-changing geographic units which hamper time series analysis.

Administrative units such as cities and towns are very unstable over time, and even at the county level there have been changes over time. The development of the census tract program over time has alleviated some of those issues, but they are often too large in terms of geographic area and population to be useful for many purposes. Over the years, we have painstakingly migrated the historical census data with each decade to the latest block group boundaries. Conceptually, it is a simple matter of allocating historical block and block group level data to the new census blocks, and reaggregating to the block group level.

While this seems quite simple for one-dimensional variables (population, households), in practice we must rely on guideposts from both periods and from alternate sources to accurately disaggregate the data from the old boundaries then reaggregate it on the new. An example of this is to carefully compare the age composition of the dwelling units between the two geography sets.

For one-dimensional variables, the final step is to convert the results to integers. As an aside, our statistical side would prefer to leave the data as is, but most users are strangely uncomfortable with the concept of 3.17348 people in a block group.

For two-dimensional tables (such as age by sex), and multidimensional tables, the techniques are much more complex. Iterative proportional fitting (IPF) techniques are generally used, but these are computationally intensive and often fail to reach convergence, especially when dealing with tables beyond two dimensions. While we make use of IPF techniques, we prefer to use maximum entropy models which are one-pass solutions that force the values of a matrix to sum to their target marginal totals while minimizing the disruption of the structural integrity of the relationships between the matrix cells. This avoids a common IPF problem that emerges because the techniques lack memory and may make repeated adjustments to cell elements that lead to distorted but stabilized results.

These multi-dimensional matrix techniques work in floating-point arithmetic, so an age by sex by race table will be in fractional people. Simply rounding these values to integers will result in sparse tables which will once again not sum to the target totals.

Here is where the geographic experience comes in, along with some fuzzy logic thinking. Block groups are nested within census tracts, census tracts within counties, and counties within states. Since no state boundary changes have occurring in quite some time, the individual cells of the state level matrices will be integers, and by successively working up and down the hierarchy tree, it is possible to reasonably allocate down to the block group level. An issue that emerges is the scale gap between the census tract and county levels. In some cases, that gap can be too large to overcome – Los Angeles County, for example. Amalgamations of census tracts can be used here where the boundaries have been consistent from one census to the next.

Lessons from Canada

The census of Canada has long utilized rather disruptive privacy shields in the release of data, by using a random-rounding technique to the nearest five. For most geographic levels, all numbers end in 0 or 5 and, if the total population is under a minimum threshold, no detailed data is presented at all. The results at the dissemination area level (roughly equivalent to a block group) are, lacking a better term here, lumpy. The distribution of population by age will not equal the total, and the cross-tabulation of age by sex will not equal either major dimension.

The techniques AGS has developed and used in the United States have been demonstrated to work even in the extreme conditions imposed by Statistics Canada and result in data which is both internally and hierarchically consistent. Our extensive experience with decades of census data in each country puts AGS in a unique position to properly curate the 2020 census data.

Our 2020 Census Approach

The additional complexity of the privacy budget concept presents additional challenges, in that even the base population counts at a census block level have been modified, sometimes even to the point of statistical impossibility.

Our approach to making the 2020 census redistricting release usable includes –

- Using multiple levels of hierarchically organized geographic units to harmonize the data from the block level to the known state totals by utilizing census tracts, counties, and stable sub-county temporary areas
- Adding external data from the ACS public use micro samples to develop maximum likelihood models of sub-tables (such as households by size)
- Using available historical data (ACS, census) that refines the micro-distribution of certain variables such as vacant dwellings and household size.
- Using the spatial relationships between census geographies (adjacency and distance measures) to shift specific variables between units within parent geographies in an intelligent manner.

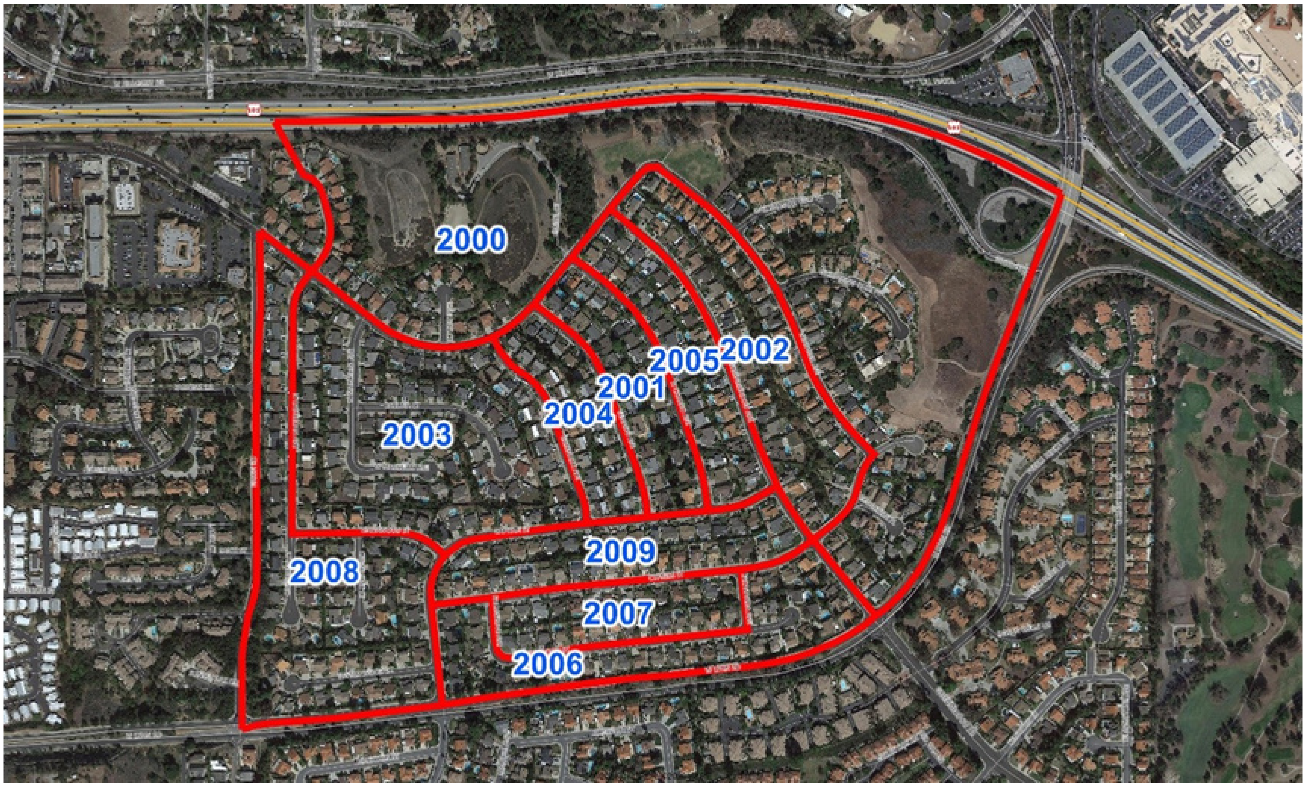
Most users of census data rarely look at individual block group data, but rather focus on trade area aggregations (radius and drive time areas) or standard geographic aggregates such as ZIP codes. Many of those users will not be aware of the issues that we address here, since many of these problems are resolved with larger geographic areas. That said, some will notice the problems when individual blocks or block groups are mapped, which, in our earlier termite analogy, demonstrates the structural damage that otherwise lies hidden below the surface.

Our considerable effort here will not result in more accurate data, as it is impossible to know the actual values. That said, it will be both internally and spatially consistent, and will effectively present the maximum likelihood solution. When used in combination with the ACS over the coming decade, it will be much more usable as a geographic base.

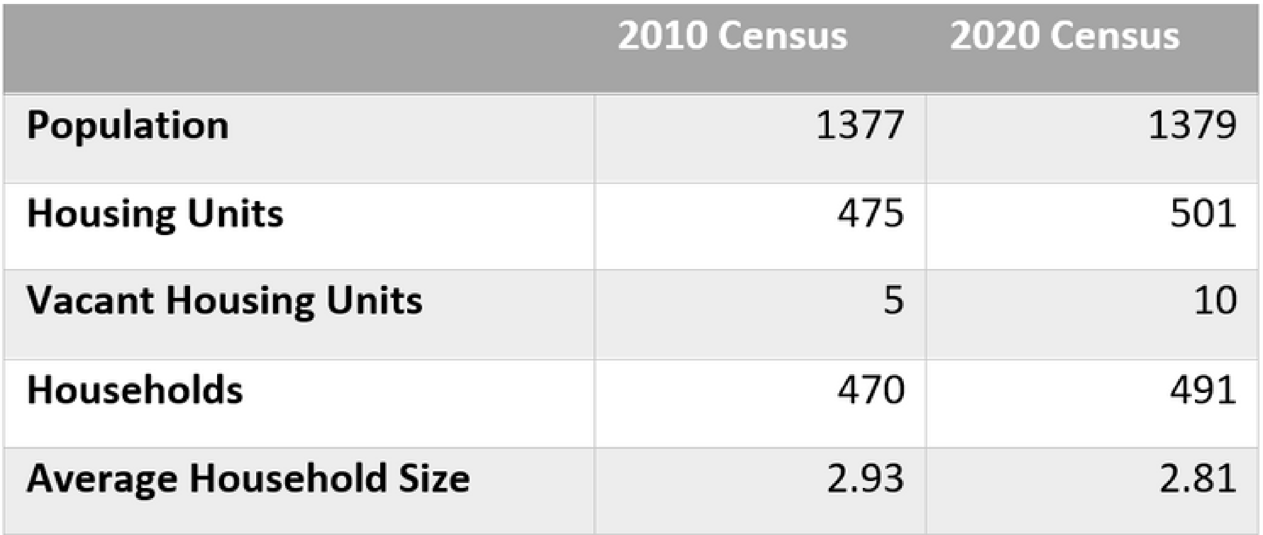
Deep Dive into 2020 Census Block Data

To further help users understand how the privacy budget has affected small area data, we decided to deep dive into a specific area to see how it looks on the ground. We chose a block group which we know that has not changed over the past ten years, located in a well-established part of Thousand Oaks, California, where our headquarters is. This is not an "outlier", and it is important to note that we found similar patterns in nearly all block groups nationwide.

The block group 061110059092 (2010) was not redefined, although the unpopulated blocks along the freeway have been merged into block 2000. For convenience, we will label them only using the 2020 numbers, as the block numbering has changed drastically. The ten blocks appear below:



It is largely a residential neighborhood, built in the 1970's, with open space along the freeway that includes an equestrian center. At a summary level, the block group has changed little over ten years. The number of homes has grown slightly with infill development, and the average household size has decreased slightly over time (as it has generally).



	2010 Census	2020 Census
Population	1377	1379
Housing Units	475	501
Vacant Housing Units	5	10
Households	470	491
Average Household Size	2.93	2.81

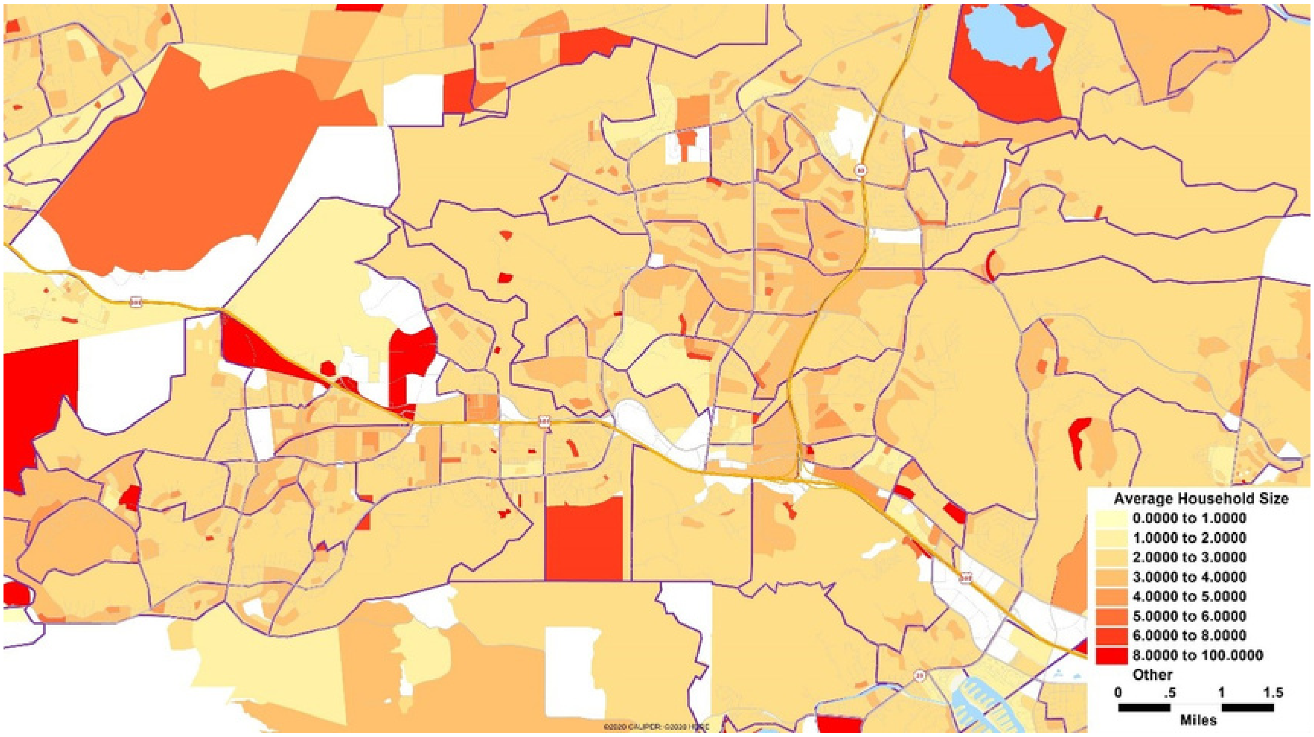
At the block level, the results are much more dramatic. The number of vacant dwellings in the block group doubled from 5 to 10, and yet all 10 are located in a single block (2004) which does not appear to be materially different than the rest. Indeed, we believe that it has vacant dwellings "borrowed" from an adjacent block group!

Further, its population increased significantly, so the average household size jumped from 2.85 (average) to 6.40. The table at the end of the article contains the data for the ten census blocks.

While a household size of 6+ persons does occur in the United States about 5% of the time, this is very abnormal in an established, upper middle-income neighborhood.

Indeed, if we map block group boundaries and display the average household size, a clear pattern emerges - almost all block groups have a single block which stands out as having a large household size (orange and red on the map below).

On closer examination, we generally find that the percentage of dwelling units vacant is substantially higher than in adjacent blocks in almost all cases.



Since AGS does its demographic modeling at the census block level, this poses particular challenges because only the total dwelling units and population in group quarters are stated to be correct at the block level. Everything else has been manipulated, and even at the block group level, there are significant anomalies.

Our approach to resolving this includes what we refer to as “balancing”, meaning that the entirety of the geographic hierarchy is utilized. State totals (stated by the census as being correct) are used to balance the county numbers, which in turn balance census tracts, block groups, and finally blocks. The outcomes are that the resulting block estimates are well constrained, and do not generally include a single block which looks nothing like its neighbors.

From an internal modeling perspective, this will yield much better results moving forward and avoid using trending on non-comparable datasets. While we can’t know what the actual census results were, we are convinced that the resulting database is likely a more accurate rendition of those results than those which have been published.

If you are interested in learning more about how we cleaned up the census data, please drop us a line. We will be happy to talk in detail about the methodology, results, and provide you with the datasets for comparison purposes.

		2020	2020
Block 2000	**2010**	**Pub**	**Rev**
Population	281	236	278
Dwelling Units	93	94	94
Vacant Units	0	0	0
Households	93	94	94
Average Household Size	3.02	2.51	2.96

		2020	2020
Block 2001	**2010**	**Pub**	**Rev**
Population	68	76	69
Dwelling Units	29	29	29
Vacant Units	1	0	0
Households	28	29	29
Average Household Size	2.43	2.62	2.38

		2020	2020
Block 2002	**2010**	**Pub**	**Rev**
Population	137	127	134
Dwelling Units	48	48	48
Vacant Units	0	0	0
Households	48	48	48
Average Household Size	2.85	2.65	2.79

		2020	2020
Block 2003	**2010**	**Pub**	**Rev**
Population	263	236	254
Dwelling Units	90	94	89
Vacant Units	0	0	0
Households	90	94	94
Average Household Size	2.92	2.51	2.70

		2020	2020
Block 2004	**2010**	**Pub**	**Rev**
Population	78	96	76
Dwelling Units	25	25	25
Vacant Units	0	10	0
Households	25	15	25
Average Household Size	3.12	6.40	3.04

		2020	2020
Block 2005	**2010**	**Pub**	**Rev**
Population	98	96	99
Dwelling Units	35	35	35
Vacant Units	1	0	0
Households	34	35	35
Average Household Size	2.88	2.74	2.83

		2020	2020
Block 2006	**2010**	**Pub**	**Rev**
Population	116	102	116
Dwelling Units	41	41	41
Vacant Units	2	0	1
Households	39	41	40
Average Household Size	2.97	2.49	2.90

		2020	2020
Block 2007	**2010**	**Pub**	**Rev**
Population	78	104	80
Dwelling Units	25	28	28
Vacant Units	0	0	1
Households	25	28	27
Average Household Size	3.12	3.71	2.96

		2020	2020
Block 2008	**2010**	**Pub**	**Rev**
Population	145	188	214
Dwelling Units	48	66	66
Vacant Units	1	0	0
Households	47	66	66
Average Household Size	3.09	2.85	3.24

		2020	2020
Block 2009	**2010**	**Pub**	**Rev**
Population	113	118	110
Dwelling Units	41	41	41
Vacant Units	0	0	0
Households	41	41	41
Average Household Size	2.76	2.88	2.68