Note: Shortly after this paper was completed the Census Bureau issued a new memo adding to the disclosure rules discussed here. A copy of that memo is attached.

Sorting Through the Census Bureau's Disclosure Rules for Small Areas

A Discussion Draft

By
Ed Christopher
Elaine Murakami
Federal Highway Administration
Department of Transportation

Nanda Srinivasan
Cambridge Systematics

This paper attempts to sort out several issues related to the data release rules being established for small area American Community Survey data. The analysis shows that small area data at the tract level and below might not exist in any reasonable or usable fashion. Throughout this analysis, several technical questions regarding the data release rules are raised. At the end is a list of questions which the Census Bureau's Disclosure Review Board needs to clarify. To assure that the American Community Survey will produce data that is practical and usable, it is critical for the Census Bureau staff and the data users to work together. Without this cooperation the success and utility of the American Community Survey is in serious jeopardy.

December 2, 2005

Small Area Data

Exhibit 1: ACS Tabulation Geographies

	Areas	
1 year data	**Tabulation areas over 65,000 people**	**Counties**
		Cities
		PUMAs
		MCDs
3 year data	**Tabulation areas between 20,000 and 65,000 people**	**Counties**
		Cities
		MCDs
5 year data	**Tabulation areas under 20,000 people**	**Cities**
		MCDs
		Tracts
		Block Groups
		TAZs

This paper will attempt to sort out several issues related to "small area data" coming from the American Community Survey (ACS). Small areas will be defined as any area falling under the Census Bureau's (CB) predefined threshold of 20,000 people. The data released for these areas is known as the "5-year" data since it will be compiled from 5 years of accumulated records. See Exhibit 1.

Under the ACS, it appears that small area data at the tract level and below may not exist in any reasonable or usable fashion for many users. Tracts represent areas of around 4,000 people. The reasons for this vary, but are generally attributed to disclosure proofing requirements and/or statistical quality rules. As will be explained, the data products in question include both CTPP like tables as well as "Standard" Census Data Tables. While disclosure and statistical quality rules apply to regular census geographies, e.g. Tracts and Block Groups (BGs), additional restrictions will be added to "custom" geographies like Traffic Analysis Zones (TAZs).

A good place to begin is with what the CB calls Standard Tables. Standard Tables are those which the CB produces as part of its regular course of business. Using Census 2000 data products as an example, Summary File 3 (SF3) represents a set of "Standard" tables while the Census Transportation Planning Package (CTPP) represents a "Special" tabulation data set. Special tabulations are paid for by the client user. In the case of the CTPP, the states and MPOs paid approximately $3 million for the 2000 CTPP tables.

Historically, the CB's Standard Tables, like SF3, have been restricted to the residence or home location. However, new to the ACS, a set of Standard Tables for workers at their place of work will be produced. CTPP aficionados know these as Part 2 tables. Accompanying these new tables, the CB's Disclosure Review Board (DRB) has established a new set of rules.

The new rules were documented in an internal CB memo dated September 13, 2005, and shared with USDOT and AASHTO at a meeting on September 16, 2005. A copy of the September 13th memo is reproduced as Attachment A and summarized in Exhibit 2.

According to our best understanding of the September 13th memo, there are two tests or rules which the standard workplace tables must pass. First, for a table to be released it must contain at least 50 unweighted records. For example, if you were interested in the modes people working in a particular Block Group (BG) used to get to work, there would

have to be at least 50 workers (individual respondents) before the table would be considered for release. If there were not 50 workers prior to weighting, the table would be suppressed.

Once a table passes the "rule of 50" suppression test, each cell within the table will be subjected to a "threshold of 3" test. Under the threshold test, there must be at least 3 unweighted workers in the cell before the data for that cell could be released. If a cell does not pass this test, it will be collapsed into another cell. For example, if there were only two workers in a BG who took a 2-person carpool to work that information would not be released. Instead, the 2-person carpool information would be collapsed into another cell. As of this writing the CB has not released or suggested that it will release the collapsing order of the cells.

Exhibit 2: ACS Table Restrictions for Standard Products

	Area	Worker Minimum	
		For Tables	For Mode to work
1 year data	Places over 65K	10	No Threshold (1)
3 year data	Places 20K to 65K	30	No Threshold (1)
5 year data	Places under 20K	50	3 per mode per area, or else it will be collapsed

Source: September 13th memo from Laura Zayatz, DRB Chair to Larry McGinn, ACS Chief.
Notes: (1) Although no thresholds will be applied by the DRB rules each cell of the all tables will be subject to a statistical test of which failure will result in collapsing.

In addition to the workplace table tests, the September 13th memo also states that the "threshold of 3" test would be applied to residence based tables when they involve tables with workers and modes. Although the memo discusses only one variable, "the means of transportation to work" we are left wondering whether the same threshold restrictions will be applied to other variables for small area tabulations. As one might expect, the new DRB disclosure rules raised many more questions than the memo answered. At the end of this article is a list of questions that still need to be resolved and clarified.

What do the new rules mean to the data?

Upon receiving the September 13th memo, we proceeded to run some independent tests to determine what the effect the new rules might have on small area data. Part 2 of the CTPP was used and two urban counties, King County in Seattle, Washington and Montgomery County in Maryland were analyzed. We selected these counties because

of our familiarity with them, and because they have relatively high proportions of transit users including, bus, ferry (King Co), and subway (Montgomery Co). Before attempting to see what the impact of these rules would be on BGs or TAZs, it was decided to first examine tract level data. Since tracts tend to be 4 to 11 times larger than BGs or TAZs, tracts were thought to represent the "best case" for small area data. The detailed methodology is shown in Appendix B.

The first step of the examination was see what the effect of the "rule of 50" would be on tracts given the assumed response rates of the ACS as compared to the 1 in 8 response rates typically associated with the Long Form for urban counties.

Exhibit 3 shows the number of tracts that failed the "rule of 50". Keep in mind that this test was performed on workplace tables. Additional tests can be done on the residence based data as well as other variables, including race, income, worker earnings and travel time once the questions raised at the end of this article are resolved.

A Word about Sampling and Response Rates

Those close to the long form are used to it being touted as a roughly 1 in 6 sample of households. The actual rate is based on an area's population density with rural areas being sampled at a higher rate (approaching 1 in 2 households) than the densely populated urban centers. For urban areas the rate is assumed to be around 1 in 8 or 12.5 percent. Due to several reasons, long form sampling rates and response rates are nearly equal. This is not the case for the ACS. The sampling rate for the ACS is in the neighborhood of that of the long form but it drops to about 1 in 14 or 16 when you consider the completed surveys. For the analysis in this article, a 1 in 14 rate was used. This rate is very consistent with that

Exhibit 3: Summary of tracts that would be suppressed under the Rule of 50

	Means of Transportation to Work		
	Total Number of Tracts	Tracts Suppressed	
		ACS	Long Form
		1 in 14 7% sample	1 in 8 13% sample
Both Counties	**550**	**193**	**77**
		35%	**14%**
Montgomery County, Maryland (Washington DC Region)	**177**	**57**	**28**
		45%	**22%**
King County, Washington (Seattle Region)	**373**	**136**	**49**
		36%	**13%**

Source: 2000 CTPP data for King County in Seattle, Washington and Montgomery County in Maryland.

Note: For a description of the methodology to arrive at the number of unweighted records refer to Attachment B.

When looking at the results in Exhibit 3, it is rather shocking to see exactly how many tracts with data would be lost simply by the application of the "rule of 50." Both counties are very urban and had people working in every tract. This might not always be the case. There are some tracts around the country that do not have any workers and would not show any data regardless of which disclosure rules were used. However, in the counties analyzed, the tracts not passing the "rule of 50" test all had some data in them.

The next step in the analysis was to examine which cells would pass the "threshold of 3" test and to see if there were any differences between the two counties. While both counties are heavily urbanized there are some obvious differences in travel modes. Both have a fair amount of transit users, but King County in the Seattle region is also home to some of the largest concentrations of ferry boat commuters, and Montgomery County, which is part of the Washington DC region, has the Metro subway system.

Exhibit 4 below presents a series of tables showing the individual modes which passed and failed the "threshold of 3" test. Keep in mind that failing this test does not mean that the data will be suppressed but instead it will be collapsed with other modes.

Exhibit 4: Analysis of the individual Modes with respect to the Rule of 3
Montgomery and King Counties

	Passed Rule of 3		**Failed** Rule of 3	
	Number	Percent	Number	Percent
Drove alone	357	100.0	0	0.0
2-person carpool	341	95.5	16	4.5
3-person carpool	126	35.3	231	64.7
4-person carpool	42	11.8	315	88.2
5-or-6-person carpool	21	5.9	336	94.1
7-or-more-person carpool	23	6.4	334	93.6
Bus or trolley bus	226	63.3	131	36.7
Streetcar or trolley car	1	0.3	356	99.7
Subway or elevated	39	10.9	318	89.1
Railroad	4	1.1	353	98.9
Ferryboat	24	6.7	333	93.3
Bicycle	46	12.9	311	87.1
Walked	198	55.5	159	44.5
Taxicab	5	1.4	352	98.6
Motorcycle	15	4.2	342	95.8
Other means	40	11.2	317	88.8
Worked at home	316	88.5	41	11.5
Tracts Passing Rule of 50	357			
Total Tracts	550			

King County

	Passed Rule of 3		**Failed** Rule of 3	
	Number	Percent	Number	Percent
Drove alone	237	100.0	0	0.0
2-person carpool	228	96.2	9	3.8
3-person carpool	85	35.9	152	64.1
4-person carpool	27	11.4	210	88.6
5-or-6-person carpool	19	8.0	218	92.0
7-or-more-person carpool	21	8.9	216	91.1
Bus or trolley bus	142	59.9	95	40.1
Streetcar or trolley car	1	0.4	236	99.6
Subway or elevated	1	0.4	236	99.6
Railroad	0	0.0	237	100.0
Ferryboat	24	10.1	213	89.9
Bicycle	42	17.7	195	82.3
Walked	143	60.3	94	39.7
Taxicab	3	1.3	234	98.7
Motorcycle	15	6.3	222	93.7
Other means	30	12.7	207	87.3
Worked at home	202	85.2	35	14.8
Tracts Passing Rule of 50	237			
Total Tracts	373			

Montgomery County

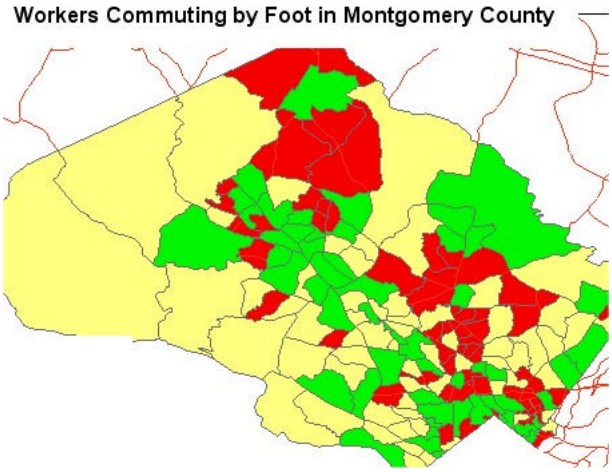
	Passed Rule of 3		**Failed** Rule of 3	
	Number	Percent	Number	Percent
Drove alone	120	100.0	0	0.0
2-person carpool	113	94.2	7	5.8
3-person carpool	41	34.2	79	65.8
4-person carpool	15	12.5	105	87.5
5-or-6-person carpool	2	1.7	118	98.3
7-or-more-person carpool	2	1.7	118	98.3
Bus or trolley bus	84	70.0	36	30.0
Streetcar or trolley car	0	0.0	120	100.0
Subway or elevated	38	31.7	82	68.3
Railroad	4	3.3	116	96.7
Ferryboat	0	0.0	120	100.0
Bicycle	4	3.3	116	96.7
Walked	55	45.8	65	54.2
Taxicab	2	1.7	118	98.3
Motorcycle	0	0.0	120	100.0
Other means	10	8.3	110	91.7
Worked at home	114	95.0	6	5.0
Tracts Passing Rule of 50	120			
Total Tracts	177			

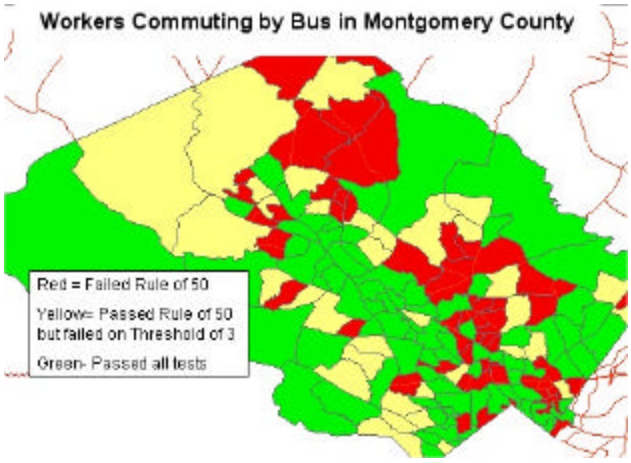
The Analysis and Commentary

Regardless of how you slice the analysis, it is clear that the "threshold of 3" test will mean a great deal of data loss even at the tract level. This loss will be compounded by the suppression caused by the "rule of 50". When looking at these results, keep in mind that this analysis was done at the tract level and that there will be an even greater loss of information when these rules are applied to smaller levels of geography like BGs and TAZs.

Just think about this for a moment. If our analysis bears out, around 45 percent of the tracts with legitimate data would be dropped into the garbage. Needless to say, there should be further testing and analysis of all of the various disclosure and other rules being applied to the Census data products. Although it was not discussed here, there are also new statistical tests that will be performed on the data products as well as some of the left over restrictions that were employed for special tabulations. Therefore, consider this a plea to all users, the CB and anyone else to help in analyzing the impacts of these rules so that useful products can be developed.

Accompanying the data loss from the two tests discussed in this paper, is a strong likelihood that different data will be available for different areas. With the collapsing rules there will be a "Swiss Cheese" effect to the data delivery. Spatially, there will be holes in the data just like Swiss cheese. To avoid this Swiss cheese approach, we believe that an aggressive research effort between DOT, the CB and the "USER" community could develop techniques for developing "synthetic" data for small areas in coordination with the various rules the CB is planning to implement.





Finally, looking at all of this from another perspective assume that you are a planner or decision-maker interested in some specific modal information like biking, walking or even transit and you need it for a neighborhood. Under the current rules, it is unlikely that you will get anything useful. Of course no one said you can't make decisions without data, or did they? For the Transportation agencies like states and MPOs, Congress in 23 U.S.C, sections 134 and 135 requires regions to have a "certified" planning and programming process and that it have a technical and analytical capability.

Unfortunately, Census data products which have been used extensively for the last 30 years have serious issues associated with them.

Have we gone too far in trying to "disclosure-proof" our data? Is this truly what the intent of Title 13 is all about?

September 13, 2005 Memo Questions

As noted earlier, the September 13th memo has raised new and additional questions about what data will be released and how it will look. Below are a series of questions that need to be answered before any further analysis can move forward. The answers to these questions will determine if other variables should be checked, other tables like resident based tables, or even possible alternatives.

1. Will the "rule of 50" and "threshold of 3" apply to all workplace tables or just the ones including the "mode used to go to work?" Consistency would suggest that it would apply to all variables and tables.

2. Will the rules in the memo be applied to the residence tables for just the "mode used to go to work" or all tables involving workers? There seems to be some disagreement over this.

3. Is there a particular reason why just the "mode used to go to work" variable was targeted by these rules? The September 13th memo only mentions the modal question.

4. Will the CB release the collapsing order before applying and implementing it? Will the order be made public or available for comment?

5. Will the collapsing schema be uniform across geographic areas or will it be different area by area? Will each tract and block group have its own collapsing schema? Will the collapsing be released or shared with the user community prior to implementation?

6. Under the "rule of 50" will the data analyst be able to identify which areas were suppressed because they failed to pass the test as opposed to those which contained true zeros? i.e. an area that had no workers.

7. Will the "rule of 50" be applied to person counts or just worker counts? The memo clearly states workers but people are asking why would the CB focus on just workers and not residents as well?

8. The September memo talks about complementary suppression towards the end. What does that mean in the context of the memo?

Attachment A: September 13th Memo

September 13, 2005

MEMORANDUM FOR Lawrence McGinn
Chief, American Community Survey Office

From: Laura V. Zayatz
Chair, Disclosure Review Board

Subject: Revised Decision on ACS Workplace Tables

The Disclosure Review Board (DRB) has reviewed and discussed your September 7, 2005 request to revise our previous decision. The Board has revised its previous decision. For workplace tables, there must be at least 10 unweighted workers in sample in a given year and a given workplace for the 1 year estimates to be shown. For workplace tables, there must be at least 30 unweighted workers in sample over the last 3 years in a given workplace for the 3 year estimates to be shown. For workplace tables, there must be at least 50 unweighted workers in sample over the last 5 years in a given workplace for the 5 year estimates to be shown. For the 1 year and 3 year estimates, there is no threshold on means of transportation (mode) for residence and workplace tables. For the 5 year estimates, there must be at least 3 unweighted workers in sample for each mode in a given place for the data to be shown for both residence and workplace tables. Otherwise the data must be collapsed or suppressed and complementary suppression must be applied.

cc: DRB (14)
Alfredo Navarro (DSSD)
Douglas Hillmer (ACSO)
Kristin Wevodau
Phillip Salopek (POP)

Attachment B: Methodology for Small Area "rule of 50" and "threshold of 3" Analysis

1. Select the County or area in question.

2. Download CTPP Table 2-003 for the appropriate geography and export to an Excel spreadsheet.

3. Unweight the data and apply the tests. Unweighting can be done different ways. Conceptually, to unweight the table you need to divide all the values by the expansion factor which is the response rate interval. In the case of the ACS test we chose a 1 in 14 interval. For the long form we used a 1 in 8 interval. Because the ACS data is assumed to represent 1 out 14 households, we could have divided all the table values by 14 and then rounded to eliminate any fractional values. However, to eliminate the "messiness" associated with rounding, we chose to apply the "rule of 50" and the unweighting, all in one step. To do this we multiplied the value one sample record was contributing to the total by 50 to establish a weighted threshold (14 X 50 = 700). Thinking about this another way, for a table to pass the "rule of 50" test it would need to have 700 or more observations, after weighting. Using this methodology one can visually look at any existing table to see if it has more than 700 observations and visually make a call if it would be suppressed.

4. The same logic and process was used for applying the "threshold of 3" rule. As a result, for a cell value to pass this test it needed to have 42 (3 X 14) or more workers for the cell to pass the threshold test.

5. This process was iteratively applied and the resultant tables in Exhibit 4 produced.

December 13, 2005

MEMORANDUM FOR **Lawrence McGinn**
Chief, American Community Survey Office

From: **Laura V. Zayatz**
Chair, Disclosure Review Board

Subject: **ACS Base Tables on Workplace and Means of Transportation Revision**

The Disclosure Review Board (DRB) has reviewed and discussed your December 5, 2005 request to revise certain tables and thresholds.

For workplace tables, there must be at least 10 unweighted or 60 weighted workers in sample in a given year and a given workplace for the 1 year estimates to be shown. For workplace tables, there must be at least 30 unweighted or 180 weighted workers in sample over the last 3 years in a given workplace for the 3 year estimates to be shown. For workplace tables, there must be at least 50 unweighted or 300 weighted workers in sample over the last 5 years in a given workplace for the 5 year estimates to be shown.

For the 1 year and 3 year estimates, there is no threshold on means of transportation (mode) for residence and workplace tables. For the 5 year estimates, there is no threshold on a univariate table of means of transportation for residence and workplace tables. For the 5 year estimates where means of transportation (mode) is crossed with 1 or more other variables, there must be at least 3 unweighted workers in sample for each mode in a given place for the data to be shown for both residence and workplace tables. Otherwise the data must be collapsed or suppressed and complementary suppression must be applied.

cc: DRB (14)
Alfredo Navarro (DSSD)
Douglas Hillmer (ACSO)
Kristin Wevodau
Lisa Blumerman